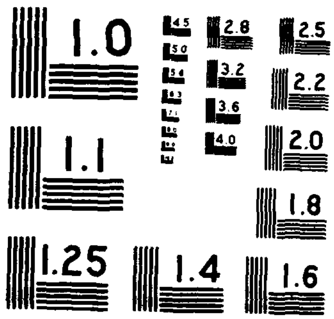
AD-A206 865 ON THE FORWARD INSTABILITY OF THE QR TRANSFORMATION(U) 1/1
CALIFORNIA UNIV BERKELEY CENTER FOR PURE AND APPLIED
MATHEMATICS J LE ET AL. JUL 88 PAM-419
UNCLASSIFIED N00014-76-C-0013 F/G 12/1 NL


1.0
4.5
5.0
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6

2

ON THE FORWARD INSTABILITY OF THE QR TRANSFORMATION

J. Le   *B. N. Parlett*

Department of Mathematics, University of California, Berkeley

May 1988

ABSTRACT

QR is the standard method for finding all the eigenvalues of a symmetric tridiagonal matrix. It produces a sequence of similar tridiagonals. It is well known that the QR transformation from T *to* $\hat{T}$ *is backward stable. That means that the computed* $\hat{T}$ *is exactly orthogonally similar to a matrix close to* T. *It is also known that the algorithm sometimes exhibits forward instability. That means that the computed* $\hat{T}$ *is not close to the exact* $\hat{T}$.

For the purpose of computing eigenvalues the property of backward stability is all that one requires. However the QR transformation has other uses and there forward stability is wanted.

This report analyzes the forward instability and shows that it occurs only when the shift causes premature deflation. We show that forward stability is governed by the behavior, in exact arithmetic, of a pair of variables and we establish tight upper and lower bounds on their derivatives with respect to change in the shift parameter.

The authors gratefully acknowledge partial support from ONR under contract N00014-76-C-0013.

89 4 18 054

(1)

1. Summary and Notation

1.1 Introduction

This study is in the area of matrix eigenvalue computations.

The *Householder-QR* algorithm has become the standard method for diagonalizing a symmetric matrix. First the matrix is reduced to tridiagonal form T by a technique introduced by *A. Householder* in 1958. Next the tridiagonal matrix T is diagonalized by successive applications of the QR transformation with shifts. Moreover it is well known ~~(see [Wilkinson, chap. 3])~~ that the QR transformation is backward stable. That means that the computed transform is exactly orthogonally similar to a matrix close to the old one. It is also known to the experts ~~(see [Wilkinson, 1958]~~, [Golub & Kahan] and [Stewart, 1970]) that the QR transformation sometimes exhibits forward instability. That means that the computed output is far from the result obtained with exact arithmetic. For the purpose of computing eigenvalues and eigenvectors the property of backward stability is all that one requires. However the QR transformation has other uses and in some cases forward stability is desirable. 2

This report analyzes the forward instability of QR and shows that it occurs only when the shift is very close to eigenvalues with a special property. For some matrices there may be no eigenvalues with this property and in such cases the algorithm is forward stable for any value of the shift. The study was prompted by the discovery that the dominant factor determining instability is the sensitivity of several key quantities to small changes in the shift parameter. This sensitivity dominates the effect of perturbation in all the other variables. (KR) (

The technical contributions of this study are:

1) The observation that instability is equivalent to premature deflation that occurs when the shift is almost an eigenvalue of several consecutive submatrices;
2) Establishing the central role of a certain pair of variables associated with each plane rotation used in the algorithm. It is the evolution of the norm of this 2-vector that governs the accuracy of the computed angles.
3) Bounds on the derivatives of the cosines of the rotation angles and other key quantities.
4) Upper and lower bounds on the last components of normalized eigenvectors in terms of eigenvalues.

1.2 Organization of this study

This first section gives the summary of this study. Along with it, the notation conventions to be used in the discussions is presented.

In **section 2**, the QR transformation is defined and several examples are exhibited to show that sometimes the QR transformation on T is forward stable and sometimes it is not. Also in **section 2**, several needed spectral properties of T are described. An useful result is **Theorem 2.3** that gives upper and lower bounds on the last element of a normalized eigenvector of T.

In **section 3**, the implementation of the QR transformation on T is discussed. In the process, the important intermediate quantities are introduced along with the relations among them.

In **section 4**, these quantities are analyzed in terms of the shift σ. The main results of this effort are 1) the derivatives of π_k, c_{k+1} (to be defined in (3.1.2)) will attain their extrema at the eigenvalues of T_k; 2) if some eigenvalue $\lambda^{(k)}$ of T_k is very close to some eigenvalue $\lambda^{(k-1)}$ of T_{k-1}, the absolute values of $d\pi_k/d\sigma$, $dc_k/d\sigma$ for σ in the vicinity of this special $\lambda^{(k)}$ can be large.

In **section 5**, the intermediate quantities produced in finite precision arithmetic by a particular QR implementation are formulated into a convenient matrix-vector form and the influence of the roundoff errors is represented by a tridiagonal error matrix. The important usage of this error matrix is in the *Perturbed Commutative Law* which is used to explain the forward instability of TQR.

1.3 Notation

Throughout this work we will use T to denote the real $n \times n$ symmetric tridiagonal matrix. In general, upper case Roman letters will denote matrices, lower case Bold letters will denote vectors and lower case Roman letters will denote scalars. Upper case Greek letters are used for special matrices (usually diagonal), lower case Greek letters are also scalars.

The matrix I_k denotes the $k \times k$ identity matrix. The matrix T_k denotes the $k-th$ leading principal submatrix of T. The norm $\|\cdot\|$ is the Euclidean norm. The transpose operation is denoted by $'$ (e.g. M' is read M transpose).

1.4 (More technical) Summary

The actual algorithm used to implement the QR transformation employs a sequence of plane rotations in planes $(1,2)$, $(2,3)$, $(3,4)$, ..., $(n-1, n)$ that change T_n into $\hat{T}_n$ by chasing a certain bulge in the tridiagonal form down the matrix and off the end. The proof that this process is equivalent to the formal definition of the QR transform (*namely (i)* $T - \sigma I = QR$; (ii) $\hat{T} = RQ + \sigma I$,) depends strongly on the <u>unreduced</u> property (to be defined in **section 2.2**) of T. In a finite precision environment we must anticipate a divergence of these two processes when any subdiagonal elements β_k $(2 \le k \le n)$ are small enough, except for the last one. In other words it is not surprising that small β_k, $(k < n)$ causes forward instability in QR.

Nevertheless this possible closeness to reducibility is not the only cause of instability. The purpose of this study is to elucidate exactly how forward instability can occur both with and without small β_k, $(k < n)$.

Forward instability, if it occurs, is quite dramatic. The cosines of the rotation angles, $c_1, c_2, ..., c_{n-1}$ exhibit the following behavior. Up to some $k < n-1$ the computed and the exact c_i have the same exponents which eventually diminish to $log(roundoff\ unit)/2$. For $i \ge k$ the true $|c_i|$ continues to decrease while the computed $|\bar{c}_i|$ increases in such a way that the products $|c_i \bar{c}_i| = O(\varepsilon)$. This holds until $|c_i| < \varepsilon$. As $|c_i|$ diminishes even further, $|\bar{c}_i|$ stays close to 1 unless β_i suddenly drops. An isolated vanishing of c_i $(i < k)$ does no harm.

One result of our study is that forward instability is always associated with "premature deflation". In the scenario given in the previous paragraph it happens that after rotation k the elements $(k-1, k)$, $(k, k-1)$, $(k, k+1)$, $(k+1, k)$ are all on the order of $\sqrt{\varepsilon}\, \|T\|$. The shift appears in position (k, k) and is correct to working precision. If row k and column k are deleted from the new matrix to obtain $\hat{T}^{(k)}$ then the eigenvalues of $\hat{T}^{(k)}$ will be the remaining eigenvalues of T_n to within working accuracy. If the shift is a well isolated eigenvalue of T_n then its eigenvector can be constructed from the rotation angles up to k.

The occurrence of forward instability is not connected with the presence of clusters of close-eigenvalues in T_n. It is caused by the shift being an eigenvalue of T_{k-1} and T_k (to working accuracy). It so happens that when this occurs the shift will be an eigenvalue of all the principal submatrices T_{k-1}, T_k, T_{k+1}, ..., T_n, (to working accuracy).

We have described instability in terms of the cosines $\bar{c}_i$ because they are more familiar. However a better indicator of stability is $(\pi_i^2 + c_i^2 \beta_{i+1}^2)^{1/2}$, where π_i is one of the variables that appears in TQR. See **section 5** for more details.

1.5 Application to the Lanczos Algorithm

In general the algorithm TQR (tridiagonal QR) is forward stable for all choice of shifts. However there is an important application where instability is endemic and it was the gradual realization of this uncomfortable fact that led to our study.

The Lanczos iteration produces a symmetric tridiagonal matrix to which a new row and column are added at each step. At the end of *step k* the algorithm has produced tridiagonal T_k and an extra number β_{k+1}. T_k represents the projection of some given linear operator A on a special k-dimensional subspace. These growing tridiagonals are special because, as k increases, <u>some</u> eigenvalues stabilize. In other words an eigenvalue of T_k appears to be equal (to working precision) to an eigenvalue of T_{k+1} and to an eigenvalue of T_{k+2}, and so on. These stabilized values are eigenvalues of the linear operator A on $\mathbf{R}^n$. In an implementation of the Lanczos algorithm it is convenient to get rid of these converged eigenvalues by

deflating them from T_n. At first sight this appears to be impossible because at *step* k ($k < n$) the matrix T_n is not fully known. However this deflation is possible and even occurs naturally (in exact arithmetic) when the QR algorithm is applied to T_{k+1} with the correct shift. The unknown component in position $(k+1, k+1)$ is not altered when the QR transformation is halted at *step* k. It is only necessary to delete row and column k.

What happened to us was that we did not always know the correct value k. When the QR transformation was forced to continue beyond the right place then the results were terrible. As a result, the deflation by QR failed and the resulting tridiagonals were wrong. Of course deflation had occurred earlier but we did not look for it. The process had encounted forward instability. In the spirit of knowing your enemy this investigation was launched.

2. Background Information

This section covers the definition of the QR transformation and its relation to eigenvalue deflation and eigenvector calculation for a symmetric tridiagonal matrix T. The examples in **section 2.3** show that the QR transformation on T can sometimes be violently unstable in the forward sense. For easy reference, several needed spectral properties of T are collected into **section 2.4** from [Parlett, chap.7]. **Theorem 2.3** gives upper and lower bounds on the bottom element of a normalized eigenvector.

2.1 The QR transformation

This well established procedure is described in several books; e.g. [Wilkinson, chap.8], [Stewart, chap.7], [Parlett, chap.8], [Golub & Van Loan, chap.7]. Here we will reproduce only what we need of the standard results. Our notation follows [Parlett, chap.8].

For any square complex matrix A and any scalar { σ } (called the shift) that excludes A's eigenvalues, the associated *QR transformation* $A \rightarrow \hat{A}$ is defined as follows:

i) let $A - \sigma I = QR$, *the unique unitary upper triangular decomposition with the diagonal elements of R being positive.*

ii) define $\hat{A} = RQ + \sigma I = Q^* A Q$.

One important property of the QR transformation is that both the upper Hessenberg form ($A = (a_{ij})$ *with* $a_{ij} = 0$ if $i > j+1$) and the Hermitian form ($A = (a_{ij})$ *with* $\bar{a}_{ij} = a_{ji}$) are preserved. Our concern here is with real symmetric tridiagonal matrices, $A = T$, and this form is preserved in the QR transformation since T is both upper Hessenberg and Hermitian. Only real shifts are considered in our investigations.

2.2 Eigenvalue deflation and eigenvector calculation

A well known result (see [Wilkinson, pp 469-471]) connects QR with eigenvalue deflation and eigenvector computation.

Definition 2.1: *A symmetric tridiagonal matrix T is called* <u>*unreduced*</u> *if its subdiagonal elements are nonzero.*

Remark: *When T is unreduced the QR transformation is well defined for all shifts* σ *because the first* $n-1$ *columns of* $T - \sigma I$ *are linear independent for all* σ.

Lemma 2.1: (QR **and deflation)**

Let T be <u>*unreduced*</u> *and* $\hat{T}$ *be the QR transform of T with shift* σ, *i.e.*

$$\hat{T} := Q^* T Q = RQ + \sigma I \tag{2.2.1}$$

where $T - \sigma I = QR$. *If* $\sigma = \lambda$, *an eigenvalue of* T, *then*

(1) last row of $\hat{T}$ *has the form* $(0, \ldots, 0, \lambda)$;
(2) last column of Q, *namely* $\mathbf{q}_n$, *satisfies*

$$T\mathbf{q}_n = \mathbf{q}_n \lambda. \tag{2.2.2}$$

Here Q *is orthogonal and* R *is upper triangular.*

Since $\hat{T} = \bar{T} \oplus \lambda$ where $\bar{T}$ has order one less than T, we say that λ has been deflated from T in one sweep of QR transformation. It is clear that the spectrum of $\bar{T}$ consists of the remaining eigenvalues of T. Also from **Lemma 2.1**, we see that when λ is deflated from T, its corresponding eigenvector is revealed in Q, namely its last column $\mathbf{q}_n$.

2.3 Some examples

In this subsection, we show, by example, that **Lemma 2.1** is not a reliable guide to results in finite precision computation. **Example 2.1** will show a successful deflation and **Example 2.2** will show a failure. **Example 2.3** will exhibit the success of deflation on the failed case in **Example 2.2** after two sweeps of QR have been applied. **Example 2.4** is an interesting case of success despite having a shift σ that is an exact eigenvalue of several of T's leading principal submatrices.

The data given in the following matrices have been multiplied by 10^4 for the purpose of better presentation. The transformed $\hat{T}$ here is generated by the numerical implementation of QR called **TQR**, which will be described in **section 3.2**.

Example 2.1: (the successful case)

$$T_6 = \begin{bmatrix} 6683.3333 & 14899.672 & & & & \\ 14899.672 & 33336.632 & 34.640987 & & & \\ & 34.640987 & 20.028014 & 11.832164 & & \\ & & 11.832164 & 20.001858 & 10.141851 & \\ & & & 10.141851 & 20.002287 & 7.5592896 \\ & & & & 7.5592896 & 20.002859 \end{bmatrix} \tag{2.3.1}$$

The eigenvalues of this matrix are:

$$\lambda_1 = 0, \quad \lambda_2 = 10, \quad \lambda_3 = 20, \quad \lambda_4 = 30, \quad \lambda_5 = 40, \quad \lambda_6 = 40000.$$

The shift is $\lambda_1 = 0$. The matrix $\hat{T}$ after one QR sweep is:

$$\hat{T}_6 = \begin{bmatrix} 39999.925 & 54.726511 & & & & \\ 54.726511 & 33.404823 & 8.3017268 & & & \\ & 8.3017268 & 24.730751 & 8.8065994 & & \\ & & 8.8065994 & 21.646903 & 7.2175779 & \\ & & & 7.2175779 & 20.292461 & -7.943d{-}12 \\ & & & & -7.943d{-}12 & -2.344d{-}15 \end{bmatrix}$$

The last row of $\hat{T}_6$ is negligible as we expected. For comparision, here is $\hat{T}_6$ computed by a method other than **TQR**.

$$\hat{T}_6 = \begin{bmatrix} 39999.925 & 54.726511 & & & & \\ 54.726511 & 33.404823 & 8.3017268 & & & \\ & 8.3017268 & 24.730751 & 8.8065994 & & \\ & & 8.8065994 & 21.646903 & 7.2175779 & \\ & & & 7.2175779 & 20.292461 & \mathbf{-1.113d{-}14} \\ & & & & \mathbf{-1.113d{-}14} & 9.520d{-}13 \end{bmatrix}$$

The matrix elements of these two transformed T_6 are almost identical except the bottom ones. However, they are negligible.

Example 2.2: (**The** **failed** **case**)

The matrix T is the same as the one in **Example 2.1**. The shift is λ_6. The matrix $\hat{T}$ after one QR sweep is:

$$\hat{T}_6^{(1)} = \begin{bmatrix} 19.989995 & 14.142133 & & & & \\ 14.142133 & 20.003002 & 11.832160 & & & \\ & 11.832160 & 20.001858 & 10.141851 & & \\ & & 10.141851 & 20.002287 & 7.5593584 & \\ & & & 7.5593584 & 20.730517 & \mathbf{-170.56153} \\ & & & & \mathbf{-170.56153} & 39999.272 \end{bmatrix} \quad (2.3.2)$$

The last subdiagonal element is not negligible. For comparison, here is $\hat{T}_6$ computed by a method other than **TQR**.

$$\hat{T}_6 = \begin{bmatrix} 19.989995 & 14.142133 & & & & \\ 14.142133 & 20.003002 & 11.832160 & & & \\ & 11.832160 & 20.001858 & 10.141851 & & \\ & & 10.141851 & 20.002287 & 7.5592896 & \\ & & & 7.5592896 & 20.002859 & \mathbf{-1.608d{-}13} \\ & & & & \mathbf{-1.608d{-}13} & 40000.000 \end{bmatrix} \quad (2.3.3)$$

Examples 2.1, 2.2 have shown that the transformation with $\sigma = \lambda_1$ is stable and the one with $\sigma = \lambda_6$ is unstable. Examination of the eigenvalues of the leading principal submatrices of T_6 (see *table 2.3.1* in **Appendix A**) reveals that λ_6 matched the biggest eigenvalues of T_3, T_4, and T_5 to almost full working precision. On the other hand λ_1 is not close to any eigenvalues of T_3, T_4 and T_5.

Example 2.3:

The tridiagonal matrix T_6 is the same as that in **Example 2.2**. We applied the QR transformation once more to the " $\hat{T}_6^{(1)}$ " exhibited in **Example 2.2** keeping the same shift $\lambda_6 = 40000$. The resulting matrix is:

$$\hat{T}_6^{(2)} = \begin{bmatrix} 19.979990 & 14.142125 & & & & \\ 14.142125 & 20.006003 & 11.832161 & & & \\ & 11.832161 & 20.003716 & 10.141851 & & \\ & & 10.141851 & 20.004574 & 7.5592897 & \\ & & & 7.5592897 & 20.005717 & \mathbf{8.425d{-}15} \\ & & & & \mathbf{8.425d{-}15} & 40000.000 \end{bmatrix} \quad (2.3.4)$$

The last subdiagonal element is now negligible.

Example 2.3 has shown that, in one case at least, **TQR** will take two sweeps to get the deflated $\hat{T}_6$. Examination of the eigenvalues of the leading principal submatrices of $\hat{T}_6^{(1)}$ obtained by first QR sweep with $\sigma = \lambda_6$ reveals that the first QR sweep does no more than destroy the closeness of the eigenvalues of T_6's leading principal submatrices, (see *table 2.3.2* in **Appendix A**).

Example 2.4: (successful deflation in an interesting case)

The matrix T in this example is the well known *second difference matrix*. The data are the original ones.

$$T_5 = \begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}$$

The eigenvalues of this matrix are:

$\lambda_1 = -2 - \sqrt{3}, \ \lambda_2 = -3, \ \lambda_3 = -2, \ \lambda_4 = -1, \ \lambda_5 = -2 + \sqrt{3}.$

The shift is $\lambda_3 = -2$. The matrix $\hat{T}_5$ after one QR sweep is:

$$\hat{T}_5 = \begin{bmatrix} -2.0000000 & 1.4142136 & & & \\ 1.4142136 & -2.0000000 & 0.70710678 & & \\ & 0.70710678 & -2.0000000 & 1.2247449 & \\ & & 1.2247449 & -2.0000000 & \mathbf{0.000000} \\ & & & \mathbf{0.0000000} & -2.00000000 \end{bmatrix}$$

One can verify that λ_3 is also an eigenvalue of the first and the third leading principal submatrices of T_5. This example shows that even if the shift is an eigenvalue of some of the leading principal submatrices, nevertheless QR deflates T in one sweep.

2.4 Spectral properties of a symmetric tridiagonal matrix T

We give here several results that we need later. They are applications of **Cauchy's Interlace Theorem**, see [Cauchy, vol. 2].

Theorem 2.1: *If T is unreduced then*

(1) the eigenvalues of T are distinct;
(2) the eigenvalues of T's consecutive leading principal submatrices interlace with each other.

For proof, see [Parlett, sec. 7-7, sec. 7-10], [Cao, chap. 2].

Definition 2.2: $spread(T) = \lambda_{max}(T) - \lambda_{min}(T)$.

Theorem 2.2: *If T is unreduced, then the subdiagonal element β_k satisfies the following inequality:*

if $n > 2$, $\quad |\beta_k| < spread(T)/2, \qquad$ for $k = 2, ..., n$;
if $n = 2$, $\quad |\beta_k| \leq spread(T)/2.$

Definition 2.3: *Let*

$$\chi_k(\sigma) := \det(T_k - \sigma I_k); \tag{2.4.1}$$

$$\delta_k := \chi_k(\sigma) / \chi_{k-1}(\sigma). \tag{2.4.2}$$

where T_k is the k-th leading principal submatrix of T.

Since χ_k is not monic it differs from the characteristic polynomial of T_k by a factor $(-1)^k$.

Notation 2.1: *We denote the eigenvalues of T by λ_i and label them so that*

$$\lambda_1 < \lambda_2 < \cdots < \lambda_n.$$

Notation 2.2: *Here we denote the bottom element of the normalized eigenvector of λ_i by ω_i.*

In our applications we need $1/\omega_i$ so we present our results in that form.

Lemma 2.2: *If T is unreduced, then*

$$\frac{1}{\omega_i^2} = -\frac{\chi_n'(\lambda_i)}{\chi_{n-1}(\lambda_i)}. \tag{2.4.3}$$

For proof, see [Parlett, chap.7].

Theorem 2.3: *Suppose T is unreduced. Let us denote the eigenvalues of T_{n-1} by μ_i so that*

$$\mu_1 < \mu_2 < \cdots < \mu_{n-1}$$

then

$$\frac{1}{\omega_i^2} > \begin{cases} \dfrac{\lambda_2 - \lambda_1}{\mu_1 - \lambda_1}, & i = 1; \\[2ex] \dfrac{\lambda_i - \lambda_{i-1}}{\lambda_i - \mu_{i-1}} \dfrac{\lambda_{i+1} - \lambda_i}{\mu_i - \lambda_i}, & i \neq 1, n; \\[2ex] \dfrac{\lambda_n - \lambda_{n-1}}{\lambda_n - \mu_{n-1}}, & i = n; \end{cases} \tag{2.4.4}$$

and

$$\frac{1}{\omega_i^2} < \begin{cases} \dfrac{\lambda_n - \lambda_1}{\mu_1 - \lambda_1}, & i = 1; \\[2ex] \dfrac{\lambda_i - \lambda_1}{\lambda_i - \mu_{i-1}} \dfrac{\lambda_n - \lambda_i}{\mu_i - \lambda_i}, & i \neq 1, n; \\[2ex] \dfrac{\lambda_n - \lambda_1}{\lambda_n - \mu_{n-1}}, & i = n. \end{cases} \tag{2.4.5}$$

Proof: $$\frac{1}{\omega_i^2} = -\frac{\chi_n'(\lambda_i)}{\chi_{n-1}(\lambda_i)}, \qquad \text{by (2.4.3).}$$

Since

$$\chi_n(\sigma) = \prod_{j=1}^{n} (\lambda_j - \sigma), \qquad \chi_{n-1}(\sigma) = \prod_{j=1}^{n-1} (\mu_j - \sigma),$$

then

$$\chi_n'(\lambda_i) = -\prod_{j=1, j\neq i}^{n}(\lambda_j - \lambda_i), \quad \chi_{n-1}(\lambda_i) = \prod_{j=1}^{n-1}(\mu_j - \lambda_i).$$

And

$$\frac{1}{\omega_i^2} = \frac{\prod_{j=1, j\neq i}^{n}(\lambda_j - \lambda_i)}{\prod_{j=1}^{n-1}(\mu_j - \lambda_i)}, \tag{2.4.6}$$

$$= \begin{cases} \prod_{j=1}^{n-1}\frac{\lambda_{j+1}-\lambda_1}{\mu_j - \lambda_1}, & i = 1; \\ \prod_{j=1}^{i-1}\frac{\lambda_i - \lambda_j}{\lambda_i - \mu_j}\prod_{j=i}^{n-1}\frac{\lambda_{j+1}-\lambda_i}{\mu_j - \lambda_i}, & i \neq 1, n; \\ \prod_{j=1}^{n-1}\frac{\lambda_n - \lambda_j}{\lambda_n - \mu_j}, & i = n. \end{cases} \tag{2.4.7}$$

By Theorem 2.1

$$\lambda_j < \mu_j < \lambda_{j+1}, \quad j = 1, ..., n-1.$$

Therefore each factor in the products in (2.4.7) is positive and is bigger than one. That is to say the formulae in (2.4.7) satisfy:

$$\frac{1}{\omega_i^2} > \begin{cases} \frac{\lambda_2 - \lambda_1}{\mu_1 - \lambda_1}, & i = 1; \\ \frac{\lambda_i - \lambda_{i-1}}{\lambda_i - \mu_{i-1}}\frac{\lambda_{i+1}-\lambda_i}{\mu_i - \lambda_i}, & i \neq 1, n; \\ \frac{\lambda_n - \lambda_{n-1}}{\lambda_n - \mu_{n-1}}, & i = n. \end{cases}$$

The reorganization of (2.4.7) reveals that

$$\prod_{j=1}^{n-1}\frac{\lambda_{j+1}-\lambda_1}{\mu_j - \lambda_1} = \frac{\lambda_n - \lambda_1}{\mu_1 - \lambda_1}\prod_{j=2}^{n-1}\frac{\lambda_j - \lambda_1}{\mu_j - \lambda_1}, \qquad i = 1.$$

$$\prod_{j=1}^{i-1}\frac{\lambda_i - \lambda_j}{\lambda_i - \mu_j}\prod_{j=i}^{n-1}\frac{\lambda_{j+1}-\lambda_i}{\mu_j - \lambda_i}$$

$$= \frac{\lambda_i - \lambda_1}{\lambda_i - \mu_{i-1}}\left(\prod_{j=1}^{i-2}\frac{\lambda_i - \lambda_{j+1}}{\lambda_i - \mu_j}\right)\frac{\lambda_n - \lambda_i}{\mu_i - \lambda_i}\left(\prod_{j=i+1}^{n-1}\frac{\lambda_j - \lambda_i}{\mu_j - \lambda_i}\right), \quad i \neq 1, n.$$

$$\prod_{j=1}^{n-1}\frac{\lambda_n - \lambda_j}{\lambda_n - \mu_j} = \frac{\lambda_n - \lambda_1}{\lambda_n - \mu_{n-1}}\prod_{j=1}^{n-2}\frac{\lambda_n - \lambda_{j+1}}{\lambda_n - \mu_j}, \qquad i = n.$$

Since $\lambda_j < \mu_j < \lambda_{j+1}$, each factor in the above products is positive and each factor in the products on the right hand sides of equal signs is smaller than one. That is to say the above formulae satisfy:

$$\frac{1}{\omega_i^2} < \begin{cases} \dfrac{\lambda_n - \lambda_1}{\mu_1 - \lambda_1} & i = 1 \\ \dfrac{\lambda_i - \lambda_1}{\lambda_i - \mu_{i-1}} \; \dfrac{\lambda_n - \lambda_i}{\mu_i - \lambda_i} & i \neq 1, n. \\ \dfrac{\lambda_n - \lambda_1}{\lambda_n - \mu_{n-1}} & i = n \end{cases}$$

■

3. Implementation of the QR transformation

This section develops the usual implementation of the QR algorithm applied to a symmetric tridiagonal matrix T. In the process, the important intermediate quantities are introduced along with the relations among them.

Most of the material is standard, see [Wilkinson, chap.8], [Stewart, chap.7], [Parlett, chap.8] and [Golub & Van Loan, chap.7]. However all the results are needed in the next section.

In the following discussion, we assume that all the tridiagonal matrices in question are *unreduced* since otherwise the problem decouples. We also assume that the offdiagonal elements β_k ($2 \le k \le n$) of T are *positive*.

3.1 QR factorization of $T - \sigma I$

The desired QR decomposition can be carried out by pre-multiplying the tridiagonal matrix $T - \sigma I$ by a sequence of plane rotation matrices R_k ($2 \le k \le n$) defined as follows.

$$R_k = \begin{bmatrix} 1 & & & & & & \\ & \cdot & & & & & \\ & & 1 & & & & \\ & & & c_k & s_k & & \\ & & & -s_k & c_k & & \\ & & & & & 1 & \\ & & & & & & \cdot \\ & & & & & & & 1 \end{bmatrix} . \quad \leftarrow k\text{-}th \text{ row} \tag{3.1.1}$$

The duty of R_k ($2 \le k \le n$) is to annihilate the ($k, k-1$) position of the matrix on the way to an upper triangular form. The formulae in **step (k)** are important for the analysis in **section 4.**

Let

$$T - \sigma I = \begin{bmatrix} \alpha_1 - \sigma & \beta_2 & & & \\ \beta_2 & \alpha_2 - \sigma & \cdot & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \beta_n \\ & & & \beta_n & \alpha_n - \sigma \end{bmatrix} .$$

It can be shown that at **step (k)** ($k < n$):

$$R_k R_{k-1} \cdots R_2 (T - \sigma I) = \begin{bmatrix} \xi_2 & \zeta_2 & \times & & & & \\ & \cdot & \cdot & \cdot & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & \xi_k & \zeta_k & \times & \\ & & & & \pi_k & c_k \beta_{k+1} & \\ & & & & \beta_{k+1} & \alpha_{k+1} - \sigma & \cdot \\ & & & & & \cdot & \cdot \end{bmatrix} \tag{3.1.2}$$

where

$$\begin{array}{ll} \xi_k = (\pi_{k-1}^2 + \beta_k^2)^{1/2} & \zeta_k = c_{k-1} c_k \beta_k + s_k (\alpha_k - \sigma) \\ c_k = \pi_{k-1} / \xi_k & s_k = \beta_k / \xi_k \\ \pi_k = -s_k \beta_k c_{k-1} + c_k (\alpha_k - \sigma). & \end{array} \tag{3.1.3}$$

At step (n):

$$R_n R_{n-1} \cdots R_2 (T - \sigma I) = \begin{bmatrix} \xi_2 & \zeta_2 & \times & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \cdot \\ & & & \xi_n & \zeta_n \\ & & & & \pi_n \end{bmatrix} = R. \tag{3.1.4}$$

$$(T - \sigma I) = Q \begin{bmatrix} \xi_2 & \zeta_2 & \times & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \cdot \\ & & & \xi_n & \zeta_n \\ & & & & \pi_n \end{bmatrix} = QR$$

with

$$Q = R_2^t R_3^t \cdots R_n^t.$$

<u>We collect some direct consequences of this decomposition</u> that are used later.

(I):

$$\chi_1(\sigma) = \alpha_1 - \sigma = \pi_1,$$
$$\chi_k(\sigma) = \xi_2 \cdots \xi_k \pi_k, \qquad 2 \le k \le n. \tag{3.1.5}$$

(II): *If T is unreduced, then*

$$\xi_k \neq 0, \quad s_k \neq 0, \qquad 2 \le k \le n. \tag{3.1.6}$$

(III): *If T is unreduced, then*

$$\pi_k(\sigma) = 0 \quad \text{if and } \textit{only} \text{ if} \quad \chi_k(\sigma) = 0, \quad 1 \le k \le n. \tag{3.1.7}$$

In words π_k vanishes only at eigenvalues of T_k. So does c_{k+1} by its definition.

Lemma 3.1: *The detailed structure of matrix Q is:*

$$Q = R_2^t R_3^t \cdots R_n^t = \begin{bmatrix} c_1c_2 & c_1(-s_2)c_3 & c_1(-s_2)(-s_3)c_4 & \cdot & \cdot & c_1(-s_2)\ldots(-s_n) \\ s_2 & c_2c_3 & c_2(-s_3)c_4 & \cdot & \cdot & \cdot \\ & s_3 & c_3c_4 & & \cdot & \cdot \\ & & s_4 & \cdot & \cdot & \cdot \\ & & & \cdot & c_{n-1}c_n & c_{n-1}(-s_n) \\ & & & & s_n & c_n \end{bmatrix}. \tag{3.1.8}$$

3.2 The QR transformation

In this subsection, we look at the inner loop of the QR transformation. The formulae $\hat{T} = Q^tTQ = RQ + \sigma I$ in (2.2.1), R in (3.1.4) and $Q = R_2^t \cdots R_n^t$ in (3.1.8) suggest a way to transform T into $\hat{T}$ without forming Q.

First let us collect the essential quantities in the QR factorization derived above. If we define

$$s_1 = 0, \qquad c_1 = 1, \qquad \pi_1 = \alpha_1 - \sigma$$

then the necessary elements of R_k, $2 \le k \le n$, can be generated by the following loop:

For k = 2, ..., n
- $\xi_k = (\pi_{k-1}^2 + \beta_k^2)^{1/2}$
- $c_k = \pi_{k-1} / \xi_k$
- $s_k = \beta_k / \xi_k$
- $\pi_k = - s_k \beta_k c_{k-1} + c_k(\alpha_k - \sigma)$.

Now, let us look at the diagonal and subdiagonal elements of matrix $RQ + \sigma I$ to find out the formulae needed to compute the matrix elements of $\hat{T}$.

By the formulae (3.1.4), (3.1.8), $RQ + \sigma I$ can be presented as follows:

$$\begin{bmatrix} \xi_2 & \zeta_2 & \times & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \cdot \\ & & & \xi_n & \zeta_n \\ & & & & \pi_n \end{bmatrix} \begin{bmatrix} c_2 & -s_2c_3 & \cdot\ \cdot & \cdot & c_1(-s_2)\ldots(-s_n) \\ s_2 & c_2c_3 & \cdot\ \cdot & \cdot & \cdot \\ & s_3 & \cdot & \cdot & \cdot \\ & & \cdot\ \cdot & \cdot & \cdot \\ & & \cdot & c_{n-1}c_n & c_{n-1}(-s_n) \\ & & & s_n & c_n \end{bmatrix} + \sigma I.$$

If we denote the diagonal and subdiagonal elements of $\hat{T}$ by $\hat{\alpha}_k$ and $\hat{\beta}_k$ respectively, direct calculations reveal that, for $2 \le k \le n$,

$$\hat{\beta}_{k-1} = \xi_k s_{k-1},$$

$$\begin{aligned} \hat{\alpha}_{k-1} &= \xi_k c_{k-1} c_k + \zeta_k s_k + \sigma, \\ &= \xi_k c_{k-1} c_k + c_{k-1} c_k \beta_k s_k + s_k^2(\alpha_k - \sigma) + \sigma, && \zeta_k = c_{k-1}c_k\beta_k + s_k(\alpha_k - \sigma) \text{ by (3.1.3)}, \\ &= c_{k-1}\pi_{k-1} + c_{k-1}c_k\beta_k s_k - c_k^2(\alpha_k - \sigma) + \alpha_k, && \pi_{k-1} = \xi_k c_k \text{ by (3.1.3)}, \\ &= c_{k-1}\pi_{k-1} - c_k(c_k(\alpha_k - \sigma) - \beta_k s_k c_{k-1}) + \alpha_k, \\ &= c_{k-1}\pi_{k-1} - c_k\pi_k + \alpha_k. \end{aligned}$$

$$\hat{\beta}_n = \pi_n s_n,$$
$$\hat{\alpha}_n = \pi_n c_n + \sigma.$$

To organize the computation it is convenient to introduce a new quantity γ_k:

Definition 3.1:

$$\gamma_k : = \pi_k c_k, \qquad k = 1, ..., n.$$

The relation between γ_k and γ_{k-1} in terms of c_k, s_k, β_k and α_k is

$$\begin{aligned} \gamma_1 &= \pi_1 c_1 \\ \gamma_k &= c_k^2(\alpha_k - \sigma) - \gamma_{k-1} s_k^2, \qquad 2 \le k \le n. \end{aligned}$$

The above formula is derived as follows

$$\begin{aligned} \pi_k c_k &= (-s_k \beta_k c_{k-1} + c_k(\alpha_k - \sigma))\, c_k & &\text{by definition of } \pi_k \text{ in (3.1.3),} \\ &= -s_k \beta_k c_k c_{k-1} + c_k^2(\alpha_k - \sigma), & & \\ &= -s_k s_k \pi_{k-1} c_{k-1} + c_k^2(\alpha_k - \sigma), & &\text{since } c_k \beta_k = \pi_{k-1} s_k \text{ from (3.1.3).} \end{aligned}$$

Linking all the above relations together, a compact algorithm emerges. We list this algorithm in detail in (4.1.1.) and name it in this study by **TQR**. It is essentially the algorithm used in the well known EISPACK collection of routines.

Example: In the following, the computed c_k's, π_k's from **TQR** in **Example 2.2** (see **section 2.3**) are listed. The resulting $\hat{T}$ is exhibited in (2.3.2). For comparison, the correct c_k's and π_k's are also listed. The resulting $\hat{T}$ is exhibited in (2.3.3).

	computed	*correct*
k	c_k	c_k
1	+1.0000000000000*d*+00	+1.0000000000000*d*+00
2	−9.1287087206375*d*−01	−9.1287087206375*d*−01
3	+7.9096456588243*d*−04	+7.9096456594300*d*−04
4	−2.3388300212821*d*−07	−2.3408762727625*d*−07
5	−8.0658942646820*d*−07	+5.9381746504385*d*−11
6	+4.2662106420853*d*−03	−1.1223688023196*d*−14

k	π_k	π_k
1	−3.3316666666667*d*+00	−3.3316666666667*d*+00
2	+2.7399802074030*d*−06	+2.7399802074446*d*−06
3	−2.7673419903274*d*−10	−2.7697632729029*d*−10
4	−8.1803099953432*d*−10	+6.0232682537306*d*−14
5	+3.2249815432264*d*−06	−1.9058339689554*d*−17
6	−1.7056308317811*d*−02	−1.6080662764449*d*−17

3.3 Analysis of the QR transformation

Section 3.2 reveals the relations between the quantities of c_k, s_k, π_k and the matrix elements α_k, β_k in one step. This is inadequate if the analysis in terms of σ is needed. In the next lemma, we present several matrix-vector relations between all the intermediate quantities generated in the QR process and the matrix T. It is these relations which will help us understand QR more deeply. These relations also tell us the

structure of an eigenvector when σ happens to be an eigenvalue of T.

Recall the partial reduction of $T - \sigma I$ to upper triangular form as it appears at step (k). It is given in (3.1.2).

The product of the plane rotation matrices $R_2, \ldots, R_k$ satisfies:

$$R_2^t R_3^t \cdots R_k^t = \left[\begin{array}{ccccc|c} c_2 & -s_2c_3 & \cdots & \cdot & c_1(-s_2)\ldots(-s_k) & \\ s_2 & c_2c_3 & \cdots & \cdot & \cdot & \\ & s_3 & \cdot & \cdot & \cdot & \\ & & \cdots & \cdot & \cdot & \\ & & \cdot & c_{k-1}c_k & c_{k-1}(-s_k) & \\ & & & s_k & c_k & \\ \hline & & & & & I_{n-k} \end{array}\right]. \tag{3.3.1}$$

The $k-th$ column of (3.3.1) plays a crucial role in our analysis.

Definition 3.2: *We denote by* $\mathbf{y}_k$ *the following vector in* $\mathbf{R}^k$.

$$\mathbf{y}_k := \begin{bmatrix} c_1(-s_2)\cdots(-s_k) \\ c_2(-s_3)\cdots(-s_k) \\ \cdot \\ \cdot \\ c_{k-1}(-s_k) \\ c_k \end{bmatrix}. \tag{3.3.2}$$

Definition 3.3: *We denote by* $\bar{\mathbf{y}}_k$ *the following vector in* $\mathbf{R}^n$.

$$\bar{\mathbf{y}}_k := R_2^t \cdots R_k^t \mathbf{e}_k = \begin{bmatrix} \mathbf{y}_k \\ 0 \\ \cdot \\ 0 \end{bmatrix}. \tag{3.3.3}$$

Equating the $k-th$ row of (3.1.2), the following matrix-vector relations emerge.

Lemma 3.2: *If T is a symmetric tridiagonal matrix of order n and σ is an arbitrary real number, then the quantities derived up to step k in* TQR *satisfy*

$$T\bar{\mathbf{y}}_k - \bar{\mathbf{y}}_k\sigma = \pi_k\mathbf{e}_k + c_k\beta_{k+1}\mathbf{e}_{k+1}, \qquad k < n, \tag{3.3.4}$$

$$\| T\bar{\mathbf{y}}_k - \bar{\mathbf{y}}_k\sigma \|^2 = \pi_k^2 + c_k^2\beta_{k+1}^2, \qquad k < n, \tag{3.3.5}$$

$$(\bar{\mathbf{y}}_k)^t(T\bar{\mathbf{y}}_k - \sigma\bar{\mathbf{y}}_k) = \pi_k c_k = \gamma_k = (\bar{\mathbf{y}}_k)^t T\bar{\mathbf{y}}_k - \sigma, \qquad k \le n, \tag{3.3.6}$$

$$T\bar{\mathbf{y}}_n - \bar{\mathbf{y}}_n\sigma = \pi_n\mathbf{e}_n. \tag{3.3.7}$$

Proof: Equate the $k-th$ row on each side of (3.1.2) and transpose to get

$$(T - \sigma I)(R_2^t \cdots R_k^t)\mathbf{e}_k = (T - \sigma I)\bar{\mathbf{y}}_k = \pi_k\mathbf{e}_k + c_k\beta_{k+1}\mathbf{e}_{k+1}.$$

Since (3.1.2) holds for $k < n$, (3.3.4) is true for $k < n$. (3.3.5), (3.3.6) are the direct results of (3.3.4). (3.3.7) is a special case of (3.3.4) since $\beta_{n+1} = 0$ when $k = n$. ■

We use notation $\bar{\mathbf{y}}_k$ instead of $\mathbf{q}_k$ since the former is slightly different from the $k-th$ column $\mathbf{q}_k$ of matrix Q. When $k = n$, $\bar{\mathbf{y}}_n = \mathbf{q}_n$.

Lemma 3.2 used T and $\bar{\mathbf{y}}_k$. Equally important is the relation between T_k, the leading $k \times k$ submatrix of T, and $\mathbf{y}_k$.

Corollary 1 of Lemma 3.2: *For any real* σ,

$$T_k \mathbf{y}_k - \mathbf{y}_k \sigma = \pi_k \mathbf{e}_k^{(k)}, \qquad 1 \le k \le n. \tag{3.3.8}$$

$$(\mathbf{y}_k)^t (T_k \mathbf{y}_k - \sigma \mathbf{y}_k) = \pi_k c_k = \gamma_k = (\mathbf{y}_k)^t T_k \mathbf{y}_k - \sigma, \quad k \le n. \tag{3.3.9}$$

Proof: Taking the first k rows of (3.3.4) and using the notation in (3.3.2), the formula in (3.3.8) is obtained for $k < n$. When $k = n$ it is the case in (3.3.7).

(3.3.9) is the direct results of (3.3.8) and (3.3.2). ■

Corollary 2 of Lemma 3.2: *If* σ *is an eigenvalue of* T_k, *then*

$$T_k \mathbf{y}_k - \mathbf{y}_k \sigma = 0, \tag{3.3.10}$$

$$T \bar{\mathbf{y}}_k - \bar{\mathbf{y}}_k \sigma = c_k \beta_{k+1} \mathbf{e}_{k+1}, \qquad \| T \bar{\mathbf{y}}_k - \bar{\mathbf{y}}_k \sigma \| = | c_k \beta_{k+1} |, \tag{3.3.11}$$

$$\sigma = (\mathbf{y}_k)^t T_k \mathbf{y}_k = (\bar{\mathbf{y}}_k)^t T \bar{\mathbf{y}}_k. \tag{3.3.12}$$

Direct consequences of (3.1.7), (3.3.8), (3.3.4), (3.3.5), (3.3.6) and (3.3.9).

Definition 3.4: *We denote by* $\mathbf{t}_k$ *the following vector in* $\mathbf{R}^2$.

$$\mathbf{t}_k := \begin{bmatrix} \pi_k \\ c_k \beta_{k+1} \end{bmatrix} \tag{3.3.13}$$

4. Properties of π_k, c_k and $\mathbf{t}_k$

The main results are (4.1.12) and the tight bounds on the derivatives of π_k, c_k and $\|\mathbf{t}_k\|$ with respect to σ. See **Theorems 4.1-4.5**.

4.1 Basic properties.

We label the eigenvalues of T_k so that

$$\lambda_1^{(k)} < \lambda_2^{(k)} < \cdots < \lambda_k^{(k)}.$$

For easy reference, **TQR** and relations (3.1.7), (3.3.8), (3.3.2) are restated here.

TQR: (4.1.1)

$$s_1 = 0, \qquad c_1 = 1, \qquad \pi_1 = \alpha_1 - \sigma, \qquad \gamma_1 = \pi_1 c_1 = \pi_1$$

for k = 2, ... , n do

$$\begin{aligned}
&\xi_k = (\pi_{k-1}^2 + \beta_k^2)^{1/2} \\
&s_k = \beta_k / \xi_k \\
&c_k = \pi_{k-1} / \xi_k \\
&\pi_k = -s_k \beta_k c_{k-1} + c_k(\alpha_k - \sigma) \\
&\gamma_k = c_k^2(\alpha_k - \sigma) - s_k^2 \gamma_{k-1} \\
&\hat{\alpha}_{k-1} = \gamma_{k-1} - \gamma_k + \alpha_k \\
&\hat{\beta}_{k-1} = \xi_k s_{k-1}
\end{aligned}$$

$$\hat{\beta}_n = \pi_n s_n, \qquad \hat{\alpha}_n = \gamma_n + \sigma.$$

Relation (3.1.7): *If T is unreduced, then*

$$\pi_k(\sigma) = 0 \quad \text{if and } only \text{ if} \quad \sigma = \lambda_i^{(k)}, \quad 1 \le i \le k, \quad 1 \le k \le n. \tag{4.1.2}$$

$$c_k(\sigma) = 0 \quad \text{if and } only \text{ if} \quad \sigma = \lambda_i^{(k-1)}, \quad 1 \le i \le k-1, \quad 2 \le k \le n. \tag{4.1.3}$$

Relation (3.3.8): $$(T_k - \sigma I_k)\,\mathbf{y}_k = \pi_k\,\mathbf{e}_k^{(k)}, \qquad 1 \le k \le n. \tag{4.1.4}$$

Relation (3.3.2):

$$\mathbf{y}_k = \begin{bmatrix} c_1(-s_2)\cdots(-s_k) \\ \cdot \\ \cdot \\ c_{k-1}(-s_k) \\ c_k \end{bmatrix}, \qquad \mathbf{y}_k^t\mathbf{y}_k = 1. \tag{4.1.5}$$

<u>*Some preliminary results:*</u>

Corollary of Relation (3.3.8): *When $\sigma = \lambda_i^{(k)}$, then $\mathbf{y}_k$ in (4.1.4) is its eigenvector for T_k.*

Notation 4.1: *We denote the bottom element of the normalized eigenvector of $\lambda_i^{(k)}$ by $\omega_{i,k}$, $i = 1, 2, ..., k$, $k = 1, ..., n$.*

Corollary of Relation (3.3.2): *When $\sigma = \lambda_i^{(k)}$,*

$$\omega_{i,k} = c_k = c_k(\lambda_i^{(k)}) \neq 0 \qquad i = 1, 2, ..., k, \; k = 1, 2, ..., n. \tag{4.1.6}$$

This is the direct consequence of (4.1.5) and (4.1.3).

We now discuss the smoothness of $\pi_k(\sigma)$, $c_k(\sigma)$, for $\sigma \in (-\infty, \infty)$.

Lemma 4.1: *If T is unreduced and c_k, s_k, π_k are the functions computed by TQR (see (4.1.1)), then $c_k, s_k, \pi_k, \|\mathbf{t}_k\|$ are real analytic on* **R**.

Proof: It may be verified that

$$\pi_k^2 = \det{}^2[\,T_k - \sigma I_k\,] / \det[\,(T_{k-1} - \sigma I_{k-1})^2 + \beta_k^2 \mathbf{e}_{k-1}\mathbf{e}_{k-1}^t\,].$$

Therefore π_k^2 is a rational function of order $2k/(2k-2)$ with no poles on the real axis. Therefore π_k and $\xi_k = (\pi_{k-1}^2 + \beta_k^2)^{1/2}$ are analytic on R for $k = 2, ..., n$. Since $\xi_k > 0$, $\|\mathbf{t}_k\| > 0$ it follows that c_k, s_k and $\|\mathbf{t}_k\|$ are also analytic on R. Note that $\pi_k^2 \to \sigma^2$ as $\sigma \to \infty$.
We want to show that π_k^2 is a weighted harmonic mean of the $(\lambda_i^{(k)} - \sigma)^2$, $i = 1, ..., k$.

Premultiply both sides of (4.1.4) by $(T_k - \sigma I_k)^{-1}\pi_k^{-1}$ to find

$$\frac{\mathbf{y}_k}{\pi_k} = (T_k - \sigma I_k)^{-1}\mathbf{e}_k^{(k)}, \qquad \text{for} \quad \sigma \neq \lambda_i^{(k)}. \tag{4.1.7}$$

A consequence of the spectral factorization is

$$(\mathbf{e}_k^{(k)})^t (T_k - \sigma I_k)^{-p}\mathbf{e}_k^{(k)} = \sum_{i=1}^{k} \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^p}, \qquad \text{for} \quad \sigma \neq \lambda_i^{(k)}. \tag{4.1.8}$$

Since $\mathbf{y}_k^t\mathbf{y}_k = 1$ by (4.1.5), (4.1.7) yields

$$\frac{1}{\pi_k^2} = (\mathbf{e}_k^{(k)})^t (T_k - \sigma I_k)^{-2}\mathbf{e}_k^{(k)}. \tag{4.1.9}$$

Combine (4.1.8) and (4.1.9) to find

$$\frac{1}{\pi_k^2} = \sum_{i=1}^{k} \left(\frac{\omega_{i,k}}{\lambda_i^{(k)} - \sigma} \right)^2, \qquad \text{for} \quad \sigma \neq \lambda_i^{(k)}. \tag{4.1.10} \blacksquare$$

Applying **Lemma 4.1**, a fundamental relation between π_k, c_k and their derivatives with respect to σ is obtained. $f'(\sigma) = df(\sigma)/d\sigma$.

Lemma 4.2: *For all real* σ

$$\pi_k c_k' - c_k \pi_k' = 1. \tag{4.1.11}$$

Proof: Differentiate (4.1.4)

$$-\mathbf{y}_k + (T_k - \sigma I_k)\mathbf{y}_k' = \pi_k' \mathbf{e}_k^{(k)}.$$

Multiply by $(\mathbf{y}_k)^t$ and recall the definition of $\mathbf{y}_k$ in (4.1.5)

$$-(\mathbf{y}_k)^t\mathbf{y}_k + (\mathbf{y}_k)^t(T_k - \sigma I_k)\mathbf{y}_k' = c_k \pi_k'.$$

The result follows since $(\mathbf{y}_k)^t\mathbf{y}_k = 1$ and $(\mathbf{y}_k)^t(T_k - \sigma I_k) = \pi_k(\mathbf{e}_k^{(k)})^t$ by (4.1.4). ■

Recall that $\mathbf{t}_k := (\pi_k, c_k\beta_{k+1})^t$.

Corollary 1 of Lemma 4.2: *For any real* σ

$$\begin{aligned} &\|\mathbf{t}_k\| \, \|\mathbf{t}_k'\| \, |\sin\angle(\mathbf{t}_k, \mathbf{t}_k')| \\ &= |\mathbf{t}_k \times \mathbf{t}_k'| \\ &= |\pi_k c_k'\beta_{k+1} - \pi_k' c_k \beta_{k+1}| \\ &= |\beta_{k+1}| \qquad \textit{by } \textbf{Lemma 4.2} \\ &= \beta_{k+1}. \ ■ \end{aligned} \tag{4.1.12}$$

<u>Discussion:</u>

Relation (4.1.12) tells us that $\|\mathbf{t}_k'\|$ will be huge if $\|\mathbf{t}_k\| \, |\sin\ (\mathbf{t}_k, \mathbf{t}_k')|$ is tiny and β_{k+1} is moderate. In order to understand the behavior of $\|\mathbf{t}_k'\|$ deeply, a detailed analysis has been made on the behaviors of π_k and c_k, of which $\|\mathbf{t}_k\|$ consists.

4.2 Properties of π_k

In this subsection we investigate the behavior of $\pi_k, 2 \le k \le n$ for $\sigma \in (-\infty, \infty)$.

Corollary 2 of Lemma 4.2:

$$\pi_k'(\lambda_i^{(k)}) = -\frac{1}{c_k(\lambda_i^{(k)})} = -\frac{1}{\omega_{i,k}}, \qquad 1 \le i \le k. \tag{4.2.1}$$

Proof: Since $\pi_k(\lambda_i^{(k)}) = 0$ by (4.1.2), relation (4.1.11) at $\lambda_i^{(k)}$ becomes $c_k(\lambda_i^{(k)})\,\pi_k'(\lambda_i^{(k)}) = -1$. Since $c_k(\lambda_i^{(k)}) = \omega_{i,k} \neq 0$ by (4.1.6), (4.2.1) is obtained. ■

Corollary 3 of Lemma 4.2: *For any eigenvalue* $\lambda_i^{(k)}$ *of* T_k,

$$\pi_k(\sigma) = \pi_k'(\lambda_i^{(k)})(\sigma - \lambda_i^{(k)})\mathbf{y}_k^t(\sigma)\mathbf{y}_k(\lambda_i^{(k)}).$$

Proof: Rewrite (4.1.4) as

$$(T_k - \sigma + \sigma - \lambda_i^{(k)})\mathbf{y}_k(\lambda_i^{(k)}) = \mathbf{e}_k \pi_k(\lambda_i^{(k)}) = 0.$$

Premultiply by $\mathbf{y}_k^t(\sigma)$ to get

$$\pi_k(\sigma)c_k(\lambda_i^{(k)}) = -(\sigma - \lambda_i^{(k)})\mathbf{y}_k^t(\sigma)\mathbf{y}_k(\lambda_i^{(k)}).$$

and then apply **Corollary 2 of Lemma 4.2.**

Corollary 4 of Lemma 4.2:

$$\pi_k''(\lambda_i^{(k)}) = 0, \qquad 1 \le i \le k. \tag{4.2.2}$$

Proof: Differentiate (4.1.11): $\pi_k c_k'' - c_k \pi_k'' = 0$. Set $\sigma = \lambda_i^{(k)}$, then $\pi_k = 0$ by (4.1.2) and $c_k = \omega_{i,k} \neq 0$ by (4.1.6). ■

We now turn to the behaviour of π_k for all other values of σ.

Lemma 4.3:

$$\pi_k \pi_k'' < 0, \qquad \text{for} \quad \sigma \neq \lambda_i^{(k)}. \tag{4.2.3}$$

Proof: Differentiate both sides of (4.1.10)

$$-\frac{1}{\pi_k^3}\pi_k' = \sum_{i=1}^{k} \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^3}, \qquad \text{for} \quad \sigma \neq \lambda_i^{(k)}. \tag{4.2.4}$$

Differentiate right hand side of (4.2.4)

$$3\sum_{i=1}^{k} \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^4}. \tag{4.2.5}$$

Differentiate left hand side of (4.2.4)

$$-\frac{1}{\pi_k^3}\pi_k'' + \frac{3}{\pi_k^4}(\pi_k')^2. \tag{4.2.6}$$

Therefore for $\sigma \neq \lambda_i^{(k)}$

$$-\frac{1}{\pi_k^5}\pi_k'' = \frac{3}{\pi_k^2}\sum_{i=1}^{k} \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^4} - \frac{3}{\pi_k^6}(\pi_k')^2 \qquad \text{by (4.2.5) and (4.2.6)}$$

$$= 3\sum_{i=1}^{k} \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^2} \sum_{i=1}^{k} \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^4} - 3\left(\frac{1}{\pi_k^3}\pi_k'\right)^2 \qquad \text{using (4.1.10)}$$

$$= 3\left[\sum_{i=1}^{k}\left(\frac{\omega_{i,k}}{\lambda_i^{(k)} - \sigma}\right)^2 \sum_{i=1}^{k} \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^4} - \left(\sum_{i=1}^{k} \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^3}\right)^2\right] \qquad \text{using (4.2.4)}$$

$$\geq 0. \qquad \textit{by } \textbf{Cauchy–Schwarz Inequality.}$$

Moreover equality holds if and only if the following two vectors

$$z_1 = \left(\frac{\omega_{1,k}}{\lambda_1^{(k)} - \sigma}, \cdots, \frac{\omega_{k,k}}{\lambda_k^{(k)} - \sigma}\right)^t$$

$$z_2 = \left(\frac{\omega_{1,k}}{(\lambda_1^{(k)} - \sigma)^2}, \cdots, \frac{\omega_{k,k}}{(\lambda_k^{(k)} - \sigma)^2}\right)^t$$

are proportional. That is to say:

$$\lambda_1^{(k)} - \sigma = \cdots = \lambda_k^{(k)} - \sigma,$$

or

$$\lambda_1^{(k)} = \cdots = \lambda_k^{(k)}.$$

That contradicts the conclusion of **Theorem 2.1** in **section 2.4**. Therefore

$$-\frac{1}{\pi_k^5}\,\pi_k''(\sigma) \;>\; 0, \qquad \text{for} \quad \sigma \neq \lambda_i^{(k)}.$$

Recall from (4.1.2) that $\pi_k(\sigma) \neq 0$ for $\sigma \neq \lambda_i^{(k)}$. Therefore

$$-\,\pi_k\,\pi_k'' \;>\; 0, \qquad \text{for} \quad \sigma \neq \lambda_i^{(k)}. \qquad \blacksquare$$

Corollary of Lemma 4.3:

$$\pi_k'' \;\neq\; 0, \qquad \text{for} \quad \sigma \neq \lambda_i^{(k)}. \tag{4.2.7}$$

Lemma 4.3 shows that the algebraic function $\pi_k(\sigma)$ is like the characteristic polynomial of T_k in that it vanishes at the eigenvalues of T_k. Moreover it is alternatingly concave upward and downward in the intervals bounded by the eigenvalues of T_k.

The next result is a direct corollary of the proceeding lemmas and **Theorem 2.3** in **section 2.4**. We illustrate it in *Fig. 4.1* which shows that π_k' attains its extreme values at the $\lambda_i^{(k)}$, $i = 1, \ldots, k$.

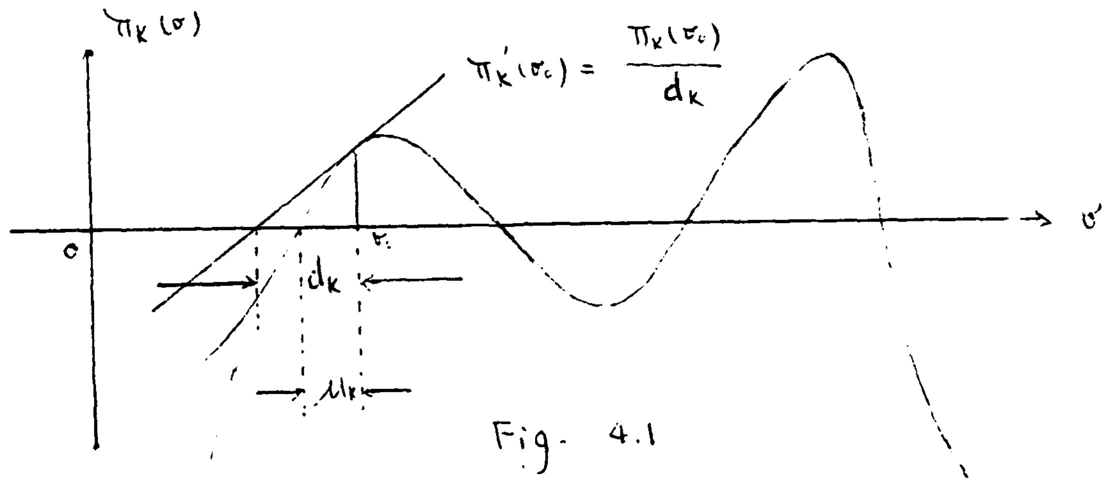


Fig. 4.1

Theorem 4.1: *The derivative of the function $\pi_k(\sigma)$ computed by* **TQR** satisfies

$$\frac{\lambda_2^{(k)} - \lambda_1^{(k)}}{\lambda_1^{(k-1)} - \lambda_1^{(k)}} \;<\; [\,\pi_k'(\lambda_1^{(k)})\,]^2 \;<\; \frac{\lambda_k^{(k)} - \lambda_1^{(k)}}{\lambda_1^{(k-1)} - \lambda_1^{(k)}}, \tag{4.2.8a}$$

$$\frac{\lambda_i^{(k)} - \lambda_{i-1}^{(k)}}{\lambda_i^{(k)} - \lambda_{i-1}^{(k-1)}}\,\frac{\lambda_{i+1}^{(k)} - \lambda_i^{(k)}}{\lambda_i^{(k-1)} - \lambda_i^{(k)}} \;<\; [\,\pi_k'(\lambda_i^{(k)})\,]^2$$

$$[\,\pi_k'(\lambda_i^{(k)})\,]^2 \;<\; \frac{\lambda_i^{(k)} - \lambda_1^{(k)}}{\lambda_i^{(k)} - \lambda_{i-1}^{(k-1)}}\,\frac{\lambda_k^{(k)} - \lambda_i^{(k)}}{\lambda_i^{(k-1)} - \lambda_i^{(k)}}, \qquad i \neq 1, k, \tag{4.2.8b}$$

$$\frac{\lambda_k^{(k)} - \lambda_{k-1}^{(k)}}{\lambda_k^{(k)} - \lambda_{k-1}^{(k-1)}} \;<\; [\,\pi_k'(\lambda_k^{(k)})\,]^2 \;<\; \frac{\lambda_k^{(k)} - \lambda_1^{(k)}}{\lambda_k^{(k)} - \lambda_{k-1}^{(k-1)}}; \tag{4.2.8c}$$

and

$$|\,\pi_k'(\sigma)\,| \le \begin{cases} |\,\pi_k'(\lambda_1^{(k)})\,|, & \sigma \in (-\infty,\, \lambda_1^{(k)}\,]; \\[2ex] \max\,[\,|\,\pi_k'(\lambda_i^{(k)})\,|,\, |\,\pi_k'(\lambda_{i+1}^{(k)})\,|\,] & \sigma \in [\,\lambda_i^{(k)},\, \lambda_{i+1}^{(k)}\,],\;\; i = 1, \ldots, k-1; \\[2ex] |\,\pi_k'(\lambda_k^{(k)})\,|, & \sigma \in [\,\lambda_k^{(k)},\, \infty\,). \end{cases}$$

Proof: By (4.2.7): $\pi_k'' \neq 0$ for $\sigma \neq \lambda_i^{(k)}$. **Apply Corollary 4 of Lemma 4.2**

$$\pi_k'' = 0 \qquad \text{for} \quad \sigma = \lambda_i^{(k)}.$$

Thus on each $[\lambda_i^{(k)}, \lambda_{i+1}^{(k)}]$ for $i = 1, ..., k-1$, π_k' has its only stationary points at the ends. For the intervals $(-\infty, \lambda_1^{(k)}]$ and $[\lambda_k^{(k)}, \infty)$ the stationary points are $\lambda_1^{(k)}$ and $\lambda_k^{(k)}$. Therefore

$$|\pi_k'(\sigma)| \leq \max[\,|\pi_k'(\lambda_i^{(k)})|,\ |\pi_k'(\lambda_{i+1}^{(k)})|\,], \qquad \text{for} \quad i = 1, ..., k-1.$$

Use (4.2.1) to find that

$$\pi_k'(\lambda_i^{(k)}) = -1/\omega_{i,k}, \qquad \text{for} \quad i = 1, ..., k,$$

Apply the bounds for $1/\omega_{i,k}^2$ in **Theorem 2.3** on T_k, the formulae in (4.2.8a), (4.2.8b) and (4.2.8c) are revealed. ■

We now give a pointwise bound on π_k' that reveals the role of the distance of σ from the spectrum of T_k and from the spectrum of T_{k-1}. Some preliminary results are restated for easy reference.

(A)	$\xi_k = (\pi_{k-1}^2 + \beta_k^2)^{1/2}$	see definition in (4.1.1)	
(B)	$c_k = \dfrac{\pi_{k-1}}{\xi_k}$	see definition in (4.1.1)	
(C)	$s_k = \dfrac{\beta_k}{\xi_k}$	see definition in (4.1.1)	
(D)	$c_k' = \dfrac{s_k^2}{\xi_k}\pi_{k-1}'$	derivative of (B)	
(E)	$\pi_k c_k' - c_k \pi_k' = 1$	for *any real* σ	(from (4.1.11))
(F)	$(T_k - \sigma I_k)\, y_k = \pi_k\, e_k^{(k)}$	for *any real* σ	(from (4.1.4))

Definition 4.1: $\mu_k := \mu_k(\sigma) = \min\limits_{1 \leq i \leq k} |\lambda_i^{(k)} - \sigma|, \qquad k = 1, 2, ..., n.$

$$(G) \qquad \mu_k^2 < \pi_k^2 \qquad\qquad \text{for} \quad \sigma \neq \lambda_i^{(k)}$$

Proof of (G): $(F)^t(F)$ yields

$$(y_k)^t (T_k - \sigma I_k)^2 y_k = \pi_k^2.$$

Then

$$\begin{aligned}
\mu_k^2 &= \min_{1 \leq i \leq k} (\sigma - \lambda_i^{(k)})^2 && \text{by **Definition 4.1**,} \\
&= \lambda_{\min}((T_k - \sigma I_k)^2) \\
&< (y_k)^t (T_k - \sigma I_k)^2 y_k && y_k \text{ cannot be an eigenvector since } \sigma \neq \lambda_i^{(k)} \\
&= \pi_k^2. && \blacksquare
\end{aligned}$$

We now present one of our key technical lemmas.

Lemma 4.4: *For any real* σ,

$$\mu_k(\sigma)\,|\pi_k'| = |\pi_k| \qquad \sigma = \lambda_i^{(k)} \tag{4.2.9a}$$

$$\mu_k(\sigma)\,|\pi_k'| < |\pi_k| \qquad \sigma \neq \lambda_i^{(k)} \tag{4.2.9b}$$

$$\mu_{k-1}(\sigma)\,|\pi_k'| < s_k^2\,|\pi_k| + (\mu_{k-1}^2(\sigma) + \beta_k^2)^{1/2} \tag{4.2.10}$$

Proof of (4.2.9a,b): When $\sigma = \lambda_i^{(k)}$ then $\pi_k = 0$ by (4.1.2), $\mu_k = 0$ by definition and (4.2.9a) holds. Now suppose that $\sigma \neq \lambda_i^{(k)}$. Premultiply (F) by $(T_k - \sigma I_k)^{-1}\pi_k^{-1}$,

$$\frac{y_k}{\pi_k} = (T_k - \sigma I_k)^{-1} e_k^{(k)}. \tag{4.2.11}$$

Differentiate (4.2.11),

$$\frac{\mathbf{y}_k' \pi_k - \mathbf{y}_k \pi_k'}{\pi_k^2} = (T_k - \sigma I_k)^{-2} \mathbf{e}_k^{(k)}. \tag{4.2.12}$$

Differentiate $\mathbf{y}_k^t \mathbf{y}_k = 1$ to obtain $\mathbf{y}_k^t \mathbf{y}_k' = 0$. Premultiply (4.2.12) by $\mathbf{y}_k^t$, getting

$$\begin{aligned} -\frac{1}{\pi_k^2}\pi_k' &= (\mathbf{y}_k)^t (T_k - \sigma I_k)^{-2} \mathbf{e}_k^{(k)} && \text{since } \mathbf{y}_k^t \mathbf{y}_k' = 0 \\ &= (\mathbf{y}_k)^t (T_k - \sigma I_k)^{-1} \mathbf{y}_k / \pi_k, && \text{by (4.2.11).} \end{aligned}$$

Thus

$$-\pi_k' = (\mathbf{y}_k)^t (T_k - \sigma I_k)^{-1} \mathbf{y}_k \pi_k,$$

and

$$|\pi_k'| = |(\mathbf{y}_k)^t (T_k - \sigma I_k)^{-1} \mathbf{y}_k| \, |\pi_k|. \tag{4.2.13}$$

Since $\sigma \neq \lambda_i^{(k)}$, $\mathbf{y}_k$ will not be an eigenvector. Therefore

$$\begin{aligned} |(\mathbf{y}_k)^t (T_k - \sigma I_k)^{-1} \mathbf{y}_k| &< \max_{\mathbf{v}^t\mathbf{v}=1} |\mathbf{v}^t (T_k - \sigma I_k)^{-1} \mathbf{v}| \\ &= \| (T_k - \sigma I_k)^{-1} \| \\ &= \frac{1}{\min\limits_{1 \le i \le k} |\lambda_i^{(k)} - \sigma|} = \frac{1}{\mu_k}. \end{aligned}$$

Put this inequality into (4.2.13) and multiply by μ_k to obtain (4.2.9b). ■

Proof of (4.2.10): When $\sigma = \lambda_i^{(k-1)}$ then $\mu_{k-1} = 0$ and (4.2.10) is immediate since $\beta_k \neq 0$ and $s_k^2 |\pi_k| \neq 0$ by (3.1.6) and (4.1.2).

Now suppose $\sigma \neq \lambda_i^{(k-1)}$.

$$\begin{aligned} |\pi_k' c_k + 1| &= |c_k' \pi_k| && \text{by (E)} \\ &= \frac{s_k^2}{\xi_k} |\pi_{k-1}' \pi_k| && \text{by (D)} \\ &< \frac{s_k^2}{\xi_k} \frac{|\pi_{k-1}|}{\mu_{k-1}} |\pi_k| && \text{by (4.2.9b)} \\ &= \frac{s_k^2 |c_k \pi_k|}{\mu_{k-1}}. && \text{by (B)} \end{aligned}$$

Hence

$$|\pi_k' c_k| < \frac{s_k^2 |c_k \pi_k|}{\mu_{k-1}} + 1.$$

Since $\sigma \neq \lambda_i^{(k-1)}$, $c_k \neq 0$ by (4.1.3). Then the above inequality can be rearranged as

$$\begin{aligned} \mu_{k-1} |\pi_k'| &< s_k^2 |\pi_k| + \frac{\mu_{k-1}}{|c_k|} \\ &= s_k^2 |\pi_k| + \mu_{k-1} \frac{\xi_k}{|\pi_{k-1}|} && \text{by (B)} \\ &= s_k^2 |\pi_k| + \mu_{k-1} \frac{(\pi_{k-1}^2 + \beta_k^2)^{1/2}}{|\pi_{k-1}|} && \text{by (A)} \\ &= s_k^2 |\pi_k| + \left(\mu_{k-1}^2 + \beta_k^2 \frac{\mu_{k-1}^2}{\pi_{k-1}^2} \right)^{1/2} \\ &< s_k^2 |\pi_k| + (\mu_{k-1}^2 + \beta_k^2)^{1/2}. && \text{by (G)} \end{aligned}$$

■

The geometric interpretation of inequality (4.2.9b) is illustrated in *Fig. 4.1*. Since the graph of π_k is concave downward or concave upward, we have

$$|\pi_k'| = \frac{|\pi_k|}{d_k} < \frac{|\pi_k|}{\mu_k}, \qquad \text{for} \quad \sigma \neq \lambda_i^{(k)}.$$

The d_k is defined on *Fig. 4.1*.

Theorem 4.2: *For any real* σ

$$|\pi_k'(\sigma)| < \min\left[\, 1 + \frac{3\,spread(T_k)}{2\,\mu_k(\sigma)},\ \frac{3}{2}\left(1 + \frac{spread(T_k)}{\mu_{k-1}(\sigma)}\right)\right]. \tag{4.2.14}$$

Proof: Recall the definition of π_k in (4.1.1): $\pi_k = c_k(\alpha_k - \sigma) - \beta_k s_k c_{k-1}$. Thus

$$\begin{aligned}
|\pi_k| &= |c_k(\alpha_k - \sigma) - \beta_k s_k c_{k-1}|, \\
&< |\alpha_k - \sigma| + |\beta_k|, && \textit{since } |c_k| \neq 1 \quad \text{for} \quad 2 \le k \le n \\
&= |\alpha_k - \lambda_i^{(k)} + \lambda_i^{(k)} - \sigma| + |\beta_k|, && \text{choose closest eigenvalue} \\
&\le |\alpha_k - \lambda_i^{(k)}| + \mu_k + spread(T_k)/2, && \textit{by } \textbf{Theorem 2.2} \\
&< spread(T_k) + \mu_k + spread(T_k)/2, && \text{since } \alpha_k \in [\lambda_1^{(k)}, \lambda_k^{(k)}] \\
&= \mu_k + 3\,spread(T_k)/2.
\end{aligned}$$

Now (4.2.9b) in **Lemma 4.4** yields

$$|\pi_k'| \le \frac{|\pi_k|}{\mu_k} < 1 + \frac{3\,spread(T_k)}{2\,\mu_k}. \tag{4.2.15}$$

Next we use the definition of π_k in a different way.

$$\begin{aligned}
|\pi_k| &< |c_k|\,|\alpha_k - \sigma| + |\beta_k|, \\
&= |c_k|\,|\alpha_k - \lambda_j^{(k-1)} + \lambda_j^{(k-1)} - \sigma| + |\beta_k|, && \text{choose closest eigenvalue;} \\
&\le |c_k|(spread(T_k) + \mu_{k-1}) + spread(T_k)/2, && \text{since } \lambda_j^{(k-1)} \in [\lambda_1^{(k)}, \lambda_k^{(k)}].
\end{aligned}$$

So

$$\begin{aligned}
s_k^2\,|\pi_k| &< s_k^2\,|c_k|(spread(T_k) + \mu_{k-1}) + s_k^2\,spread(T_k)/2 \\
&\le (spread(T_k) + \mu_{k-1})/2 + spread(T_k)/2, && \text{since } |s_k c_k| \le 1/2. \\
&= \mu_{k-1}/2 + spread(T_k).
\end{aligned} \tag{4.2.16}$$

Next we bound the term $(\mu_{k-1}^2 + \beta_k^2)^{1/2}$ on the right hand side of (4.2.10).

$$(\mu_{k-1}^2 + \beta_k^2)^{1/2} \le (\mu_{k-1}^2 + (spread(T_k)/2)^2)^{1/2} < \mu_{k-1} + spread(T_k)/2. \tag{4.2.17}$$

Add (4.2.16) and (4.2.17) to get

$$s_k^2\,|\pi_k| + (\mu_{k-1}^2 + \beta_k^2)^{1/2} < \frac{3}{2}(\mu_{k-1} + spread(T_k)). \tag{4.2.18}$$

Substitute (4.2.18) into (4.2.10) to find

$$| \pi_k' | \ < \ \frac{3}{2} (\ 1 + \frac{spread(T_k)}{\mu_{k-1}} \ ). \tag{4.2.19}$$

The combination of (4.2.15) and (4.2.19) yields (4.2.14). ■

* Note that μ_k and μ_{k-1} cannot vanish simultaneously by **Theorem 2.1 in section 2.4.**

Theorem 4.2 shows that $\pi_k'(\sigma)$ can only be huge if σ is very close to an eigenvalue of T_k and an eigenvalue of T_{k-1}. For instance when $| \pi_k' | \geq 1 / \sqrt{\varepsilon}$, (4.2.14) implies

$$\frac{\mu_k}{spread(T_k)} \ < \ \frac{3\sqrt{\varepsilon}}{2(1-\sqrt{\varepsilon})}, \qquad \frac{\mu_{k-1}}{spread(T_k)} \ < \ \frac{3\sqrt{\varepsilon}}{2-3\sqrt{\varepsilon}}.$$

4.3 Properties of c_{k+1} and s_{k+1}

In this subsection we investigate the behavior of c_{k+1}, s_{k+1}, $1 \leq k \leq n-1$ as functions of the shift σ. For easy reference, we collect the previous results that are needed.

(H)	$s_{k+1}' = -\frac{s_{k+1}c_{k+1}}{\xi_{k+1}} \pi_k'$	derivative of (C)
(I)	$c_{k+1}'' = -\frac{3\,\pi_k\,\beta_{k+1}^2}{\xi_{k+1}^5} (\pi_k')^2 + \frac{\beta_{k+1}^2}{\xi_{k+1}^3} \pi_k''$,	derivative of (D)
(J)	$\pi_k \pi_k'' < 0 \quad$ for $\quad \sigma \neq \lambda_i^{(k)}$	(from (4.2.3))
(K)	$\pi_k'(\lambda_i^{(k)}) = -\frac{1}{\omega_{i,k}}$	(from (4.2.1))

Lemma 4.5: *For* $\sigma \neq \lambda_i^{(k)}$, $i = 1, \ldots, k$,

$$c_{k+1}'' \ \neq \ 0. \tag{4.3.1}$$

Proof:

$$c_{k+1}'' = -\frac{3\,\pi_k\,\beta_{k+1}^2}{\xi_{k+1}^5} (\pi_k')^2 + \frac{\beta_{k+1}^2}{\xi_{k+1}^3} \pi_k'', \qquad \text{by (I).}$$

However

$$\begin{aligned} | c_{k+1}'' | &= \frac{3\,|\pi_k|\,\beta_{k+1}^2}{\xi_{k+1}^5} (\pi_k')^2 + \frac{\beta_{k+1}^2}{\xi_{k+1}^3} | \pi_k'' |, && \text{by (J)} \\ &\geq \frac{\beta_{k+1}^2}{\xi_{k+1}^3} | \pi_k'' |, \\ &> 0. && \text{by (J) again.} \ ■ \end{aligned}$$

Lemma 4.6:

$$c_{k+1}'' (\lambda_i^{(k)}) = 0, \qquad 1 \leq i \leq k. \tag{4.3.2}$$

Proof: Use (I) and observe that $\pi_k''(\lambda_i^{(k)}) = 0, 1 \leq i \leq k$ by **Corollary 4 of Lemma 4.2** and $\pi_k(\lambda_i^{(k)}) = 0, 1 \leq i \leq k$ by (4.1.2). ■

Lemma 4.7: $c_{k+1}'(\sigma)$ reaches its extremes at the $\lambda_i^{(k)}$ and then

$$c_{k+1}'(\lambda_i^{(k)}) = \frac{\pi_k'(\lambda_i^{(k)})}{\beta_{k+1}} = -\frac{1}{\beta_{k+1}\,\omega_{i,k}} \qquad 1 \leq i \leq k \tag{4.3.3}$$

Proof: Since $c_{k+1}' \in C^1(-\infty, \infty)$, **Lemma 4.5** and **Lemma 4.6** show that c_{k+1}' reaches its extrema at $\lambda_i^{(k)}$ for $i = 1, \ldots, k$. Therefore

$$c_{k+1}'(\lambda_i^{(k)}) = \frac{s_{k+1}^2(\lambda_i^{(k)})}{(\pi_k^2 + \beta_{k+1}^2)^{1/2}} \pi_k'(\lambda_i^{(k)}), \qquad \text{by (D) and (A)}$$

$$= \frac{1}{\beta_{k+1}} \pi_k'(\lambda_i^{(k)}), \qquad \text{by (4.1.2) and (4.1.3)}$$
$$= -\frac{1}{\beta_{k+1}\, \omega_{i,k}}. \qquad \text{by (K)} \qquad \blacksquare$$

Theorem 4.3: *The derivative of function $c_{k+1}(\sigma)$ computed by* **TQR** satisfies

$$\frac{\lambda_2^{(k)} - \lambda_1^{(k)}}{\lambda_1^{(k-1)} - \lambda_1^{(k)}} < [\beta_{k+1} c_{k+1}'(\lambda_1^{(k)})]^2 < \frac{\lambda_k^{(k)} - \lambda_1^{(k)}}{\lambda_1^{(k-1)} - \lambda_1^{(k)}}, \tag{4.3.4a}$$

$$\frac{\lambda_i^{(k)} - \lambda_{i-1}^{(k)}}{\lambda_i^{(k)} - \lambda_{i-1}^{(k-1)}} \frac{\lambda_{i+1}^{(k)} - \lambda_i^{(k)}}{\lambda_i^{(k-1)} - \lambda_i^{(k)}} < [\beta_{k+1} c_{k+1}'(\lambda_i^{(k)})]^2$$
$$[\beta_{k+1} c_{k+1}'(\lambda_i^{(k)})]^2 < \frac{\lambda_i^{(k)} - \lambda_1^{(k)}}{\lambda_i^{(k)} - \lambda_{i-1}^{(k-1)}} \frac{\lambda_k^{(k)} - \lambda_i^{(k)}}{\lambda_i^{(k-1)} - \lambda_i^{(k)}}, \qquad i \neq 1, k, \tag{4.3.4b}$$

$$\frac{\lambda_k^{(k)} - \lambda_{k-1}^{(k)}}{\lambda_k^{(k)} - \lambda_{k-1}^{(k-1)}} < [\beta_{k+1} c_{k+1}'(\lambda_k^{(k)})]^2 < \frac{\lambda_k^{(k)} - \lambda_1^{(k)}}{\lambda_k^{(k)} - \lambda_{k-1}^{(k-1)}}; \tag{4.3.4c}$$

and

$$|c_{k+1}'(\sigma)| \le \begin{cases} |c_{k+1}'(\lambda_1^{(k)})|, & \sigma \in (-\infty, \lambda_1^{(k)}]; \\ \max[\, |c_{k+1}'(\lambda_i^{(k)})|, |c_{k+1}'(\lambda_{i+1}^{(k)})| \,] & \sigma \in [\lambda_i^{(k)}, \lambda_{i+1}^{(k)}],\ i = 1, \ldots, k-1; \\ |c_{k+1}'(\lambda_k^{(k)})|, & \sigma \in [\lambda_k^{(k)}, \infty). \end{cases}$$

Direct consequences of **Lemma 4.7** and **Theorem 2.3 in section 2.4.**

<u>Discussion:</u>

Lemma 4.5 tells us that c_{k+1} is like the characteristic polynomial of T_k in that it vanishes at the eigenvalues of T_k. Moreover it is alternatingly concave upward and concave downward in the intervals divided by the eigenvalues of T_k. The direct result from this geometric property of c_{k+1} is that

$$|c_{k+1}'| = \frac{|c_{k+1}|}{d_{k+1}'} < \frac{|c_{k+1}|}{\mu_k}, \qquad \text{for} \quad \sigma \neq \lambda_i^{(k)}, \tag{4.3.5}$$

as shown in the figure *Fig. 4.2*.

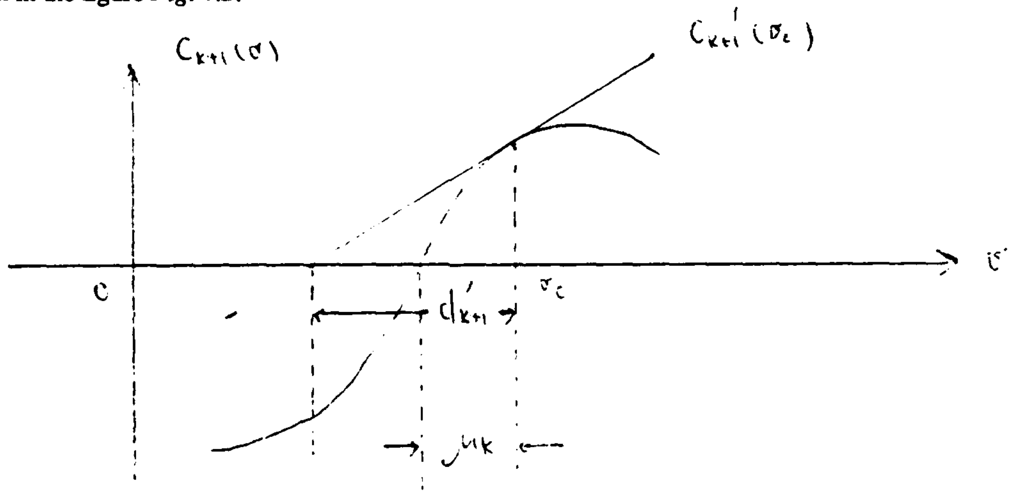


Fig. 4.2

The inequality (4.3.5) can be changed into:

$$\mu_k \mid c_{k+1}' \mid \; \le \; \mid c_{k+1} \mid \; < \; 1, \qquad \text{for } \textit{any real } \sigma.$$

However, this bound can be improved.

Lemma 4.8: *For any real* σ,

$$\mu_k \mid c_{k+1}' \mid \; < \; 1/2, \tag{4.3.6}$$

$$\mu_k \mid s_{k+1}' \mid \; < \; 1/2. \tag{4.3.7}$$

Proof: When $\sigma = \lambda_i^{(k)}$ then μ_k vanishes and the assertions are vacuously true. Now we suppose that $\sigma \ne \lambda_i^{(k)}$.

$$\begin{aligned}
\mu_k \mid c_{k+1}' \mid &= \frac{s_{k+1}^2}{\xi_{k+1}} \mu_k \mid \pi_k' \mid, && \text{by (D)} \\
&< \frac{s_{k+1}^2}{\xi_{k+1}} \mid \pi_k \mid, && \text{using (4.2.9b)} \\
&= s_{k+1}^2 \mid c_{k+1} \mid, && \text{by (B)} \\
&\le 1/2. && \text{since } \mid s_{k+1} c_{k+1} \mid \le 1/2
\end{aligned}$$

Similarly

$$\begin{aligned}
\mu_k \mid s_{k+1}' \mid &= \frac{\mid s_{k+1} c_{k+1} \mid}{\xi_{k+1}} \mu_k \mid \pi_k' \mid, && \text{by (H)} \\
&< \frac{\mid s_{k+1} c_{k+1} \mid}{\xi_{k+1}} \mid \pi_k \mid, && \text{by (4.2.9b)} \\
&= c_{k+1}^2 \mid s_{k+1} \mid, && \text{by (B)} \\
&< 1/2. && \text{since } \mid s_{k+1} c_{k+1} \mid \le 1/2 \text{ and } \mid c_{k+1} \mid < 1. \quad \blacksquare
\end{aligned}$$

As before we need to complement the proceeding result with another that involves the eigenvalues of T_{k+1}.

Lemma 4.9: *For any real* σ

$$\mu_{k+1} \mid c_{k+1}' \mid \; < \; 2 \tag{4.3.8}$$

Proof: When $\sigma = \lambda_i^{(k+1)}$ then μ_{k+1} vanishes and the assertion is vacuously true. Now we suppose that $\sigma \ne \lambda_i^{(k+1)}$. As before

$$\begin{aligned}
\mid \pi_{k+1} c_{k+1}' \mid &= \mid 1 + \pi_{k+1}' c_{k+1} \mid, && \text{by (E)} \\
&\le 1 + \mid \pi_{k+1}' \mid \mid c_{k+1} \mid, \\
&< 1 + \frac{\mid \pi_{k+1} \mid}{\mu_{k+1}} \mid c_{k+1} \mid. && \text{using (4.2.9b)}
\end{aligned}$$

Multiply each side by $\mu_{k+1} / \mid \pi_{k+1} \mid$ to get

$$\begin{aligned}
\mu_{k+1} \mid c_{k+1}' \mid &< \frac{\mu_{k+1}}{\mid \pi_{k+1} \mid} + \mid c_{k+1} \mid, \\
&< 1 + \mid c_{k+1} \mid < 2, && \text{by (G).} \quad \blacksquare
\end{aligned}$$

The above lemmas are summarized into the following theorem.

Theorem 4.4: *For any real* σ

$$\mid c_{k+1}' \mid \; < \; \min\left[\frac{1}{2\mu_k}, \frac{2}{\mu_{k+1}} \right]. \tag{4.3.9}$$

From **Theorem 4.1** and **Theorem 4.3** we see that derivatives of π_k, c_{k+1} at $\lambda_i^{(k)}$ can be huge if $|\omega_{i,k}|$ is tiny or $|\beta_{k+1}|$ is tiny or $|\omega_{i,k}\beta_{k+1}|$ is tiny.

4.4 Properties of $\mathbf{t}_k$

This subsection concentrates on $\mathbf{t}_k = (\pi_k(\sigma) + c_k(\sigma)\beta_{k+1})^t$. Since $\|\mathbf{t}_k'\| = ((\pi_k')^2 + (c_k'\beta_{k+1})^2)^{1/2}$, the results in **sections 4.2-4.3** can be applied here to bound $\|\mathbf{t}_k'\|$ as the following theorems show.

Theorem 4.5: *For any real* σ

$$\mu_k^2 \|\mathbf{t}_k'\|^2 < [\mu_k + \frac{3}{2} spread(T_k)]^2 + spread^2(T_{k+1}). \tag{4.4.1}$$

$$\mu_{k-1}^2 \|\mathbf{t}_k'\|^2 < \frac{9}{4}[\mu_{k-1} + spread(T_k)]^2 + \frac{1}{16} spread^2(T_{k+1}). \tag{4.4.2}$$

Proof: Use (4.2.15) and (4.3.8) to get

$$\mu_k^2 |\pi_k'|^2 < [\mu_k + \frac{3}{2} spread(T_k)]^2 \quad \text{and} \quad \mu_k^2 |c_k'|^2 < 4.$$

Therefore apply **Theorem 2.2** to get (4.4.1).

Use (4.2.18), (4.2.10) and (4.3.6) to get

$$\mu_{k-1}^2 |\pi_k'|^2 < \frac{9}{4}[\mu_{k-1} + spread(T_k)]^2 \quad \text{and} \quad \mu_{k-1}^2 |c_k'|^2 < \frac{1}{4.}$$

Apply **Theorem 2.2** to get (4.4.2). ■

Theorem 4.5 shows that when σ is far from the spectrum of T_k and T_{k-1}, $\|\mathbf{t}_k'\|$ has modest magnitude even when the spectra of T_k and T_{k-1} have elèments that are very close.

Remark 1: A similar result on $\|\mathbf{t}_k'\|$ to **Theorem 4.1** and **Theorem 4.3** could be derived. However it is too complicated since $|\pi_k'|$ is related to the spectra of T_k, T_{k-1} and $|c_k'|$ is related to the spectra of T_{k-1}, T_{k-2} and the magnitude of β_k. We present a simplified one.

Theorem 4.6: *The norm of vector* $\mathbf{t}_k'$ *at eigenvalues of* T_k *satisfies*

$$\frac{\lambda_2^{(k)} - \lambda_1^{(k)}}{\lambda_1^{(k-1)} - \lambda_1^{(k)}} < [\pi_k'(\lambda_1^{(k)})]^2 < \|\mathbf{t}_k'(\lambda_1^{(k)})\|,$$

$$\frac{\lambda_i^{(k)} - \lambda_{i-1}^{(k)}}{\lambda_i^{(k)} - \lambda_{i-1}^{(k-1)}} \frac{\lambda_{i+1}^{(k)} - \lambda_i^{(k)}}{\lambda_i^{(k-1)} - \lambda_i^{(k)}} < [\pi_k'(\lambda_i^{(k)})]^2 < \|\mathbf{t}_k'(\lambda_i^{(k)})\|, \quad i = 2, \ldots, k\text{-}1$$

$$\frac{\lambda_k^{(k)} - \lambda_{k-1}^{(k)}}{\lambda_k^{(k)} - \lambda_{k-1}^{(k-1)}} < [\pi_k'(\lambda_k^{(k)})]^2 < \|\mathbf{t}_k'(\lambda_k^{(k)})\|.$$

Remark 2: It will be shown in **section 5** that forward instability can appear at *step* k only if σ is very close to an eigenvalue of T_k.

5. TQR in Finite Precision Arithmetic

This section studies the relations among the computed quantities generated by **TQR** using finite precision arithmetic. A central objective is to explain the different phenomena we have observed in the examples presented in **section 2.3**. It turns out that the magnitude of the exact $\|\mathbf{t}_k'\| \, |\sin(\mathbf{t}_k, \mathbf{t}_k')|$ (defined in **Corollary 1 of Lemma 4.2**) governs the accuracy of the computed $\|\mathbf{t}_k\|$, and hence the accuracy of

the

computed π_k and c_k. From a computational point of view, the role of $\| \mathbf{t}_k' \| \, | \sin \; (\mathbf{t}_k, \mathbf{t}_k') \, |$ can be replaced by the magnitude of the exact $\| \mathbf{t}_k \|$ (to be shown in (5.2.6)). The combination of analysis in this section with **Theorem 4.5** tells us that to have forward instability, it is necessary and sufficient that σ be simultaneously close to an eigenvalue of T_k and an eigenvalue of T_{k-1} for some $k < n$.

5.1 Perturbed commutative law

We analyze *TQR* in finite precision arithmetic. If ν is an output of *TQR* in exact arithmetic then let $\bar{\nu}$ denote the corresponding output in finite precision arithmetic. In particular

Definition 5.1: *We denote by* $\bar{\mathbf{y}}_k$ *the following vector in* $\mathbf{R}^k$,

$$\bar{\mathbf{y}}_k := \begin{bmatrix} \bar{c}_1(-\bar{s}_2)\cdots(-\bar{s}_k) \\ \bar{c}_2(-\bar{s}_3)\cdots(-\bar{s}_k) \\ \cdot \\ \cdot \\ \cdot \\ \bar{c}_{k-1}(-\bar{s}_k) \\ \bar{c}_k \end{bmatrix}. \tag{5.1.1}$$

It is not the case that $\bar{\mathbf{y}}_k$ and π_k are the output of *TQR* acting on a perturbed matrix with the same shift σ. Nevertheless $\bar{\mathbf{y}}_k$ and π_k do satisfy exactly a perturbed version of the fundamental relation

$$(T_k - \sigma I_k) \mathbf{y}_k = \pi_k \mathbf{e}_k^{(k)}, \qquad k = 1, \ldots, n. \tag{5.1.2}$$

It turns out that

$$(T_k - \sigma I_k + F_k) \bar{\mathbf{y}}_k = \bar{\pi}_k \mathbf{e}_k^{(k)}, \qquad k = 1, \ldots, n. \tag{5.1.3}$$

where F_k is a tridiagonal matrix which is a small perturbation (component by component) of T_k. In fact

$$\| F_k \| < 5.6 \, spread \, \varepsilon, \qquad spread = \lambda_{\max}(T_k) - \lambda_{\min}(T_k) \tag{5.1.4}$$

provided that

$$\lambda_{\min}(T_k) \le \sigma \le \lambda_{\max}(T_k).$$

For verification of relations (5.1.3) and (5.1.4), see **section 5.3**.

The significance of (5.1.3) lies in the following *perturbed commutative law*:

Lemma 5.1: For any real σ,

$$c_k(\sigma)\beta_{k+1}\bar{\pi}_k(\sigma) - \pi_k(\sigma)\bar{c}_k(\sigma)\beta_{k+1} = \beta_{k+1} \mathbf{y}_k^t F_k \bar{\mathbf{y}}_k, \qquad 1 \le k \le n-1. \tag{5.1.5}$$

Proof: For simplicity, we drop reference to σ. Apply $\mathbf{y}_k^t$ to both sides of (5.1.3)

$$\mathbf{y}_k^t(T_k - \sigma I_k)\bar{\mathbf{y}}_k + \mathbf{y}_k^t F_k \bar{\mathbf{y}}_k = \bar{\pi}_k c_k.$$

Use (5.1.2) and multiply by β_{k+1} to get (5.1.5). ■

5.2 Forward instability of *TQR*

In **section 2.3**, we saw that *TQR* is forward unstable in **Example 2.2** and forward stable in **Example 2.4**. The interesting case is **Example 2.4**. Since the shift $\sigma = -2$ is an eigenvalue of T_3 (the 3×3 leading principal submatrix of T_5), the computed $\pi_3(-2)$ will be tiny. Obviously, it has lost all significant digits.

However the computed $\pi_4(-2)$ still preserves most of its significant digits, unlike **Example 2.2** in which the computed $\pi_k(\lambda_6)$ loses all the significant digits for $k = 4, 5$ and 6, (see table at the end of **section 3.2**). The key difference between these two examples is the lengths of residual vectors $\mathbf{t}_k(-2)$, $k = 1, \ldots, 5$ in **Example 2.4** and the lengths of residual vectors $\mathbf{t}_k(\lambda_6)$, $k = 1, \ldots, 6$ in **Example 2.2** as shown in the following table:

k	$\lVert \mathbf{t}_k(-2) \rVert$ (*Example* 2.4)	$\lVert \mathbf{t}_k(\lambda_6) \rVert$ (*Example* 2.2)
1	1.00000000*d*+00	3.64965820*d*+00
2	1.00000000*d*+00	3.16227602*d*−03
3	5.00000000*d*−01	9.35882271*d*−07
4	5.00000000*d*−01	2.37408198*d*−10
5	1.30404754*d*−33	4.48883862*d*−14
6		1.60806628*d*−17

The lengths of $\lVert \mathbf{t}_k(-2) \rVert$ never go under the magnitude of $O(\mathit{spread}\,\sqrt{\varepsilon})$ except at last step. This is what we expected for deflation to occur. However the length of $\lVert \mathbf{t}_k(\lambda_6) \rVert$ has gone under the magnitude of $O(\mathit{spread}\,\sqrt{\varepsilon})$ starting from *step* 4. The role of $\lVert \mathbf{t}_k \rVert$ is shown by the following relation.

Corollary of Lemma 5.1: *For any real* σ

$$\lVert \mathbf{t}_k(\sigma) \rVert \; \lVert \bar{\mathbf{t}}_k(\sigma) \rVert \; | \sin\angle(\mathbf{t}_k, \bar{\mathbf{t}}_k) | \;=\; \beta_{k+1} \, | \mathbf{y}_k^t F_k \bar{\mathbf{y}}_k |, \qquad k = 1, \ldots, n-1. \tag{5.2.1}$$

where $\bar{\mathbf{t}}_k(\sigma) = (\bar{\pi}_k, \bar{c}_k \beta_{k+1})^t$

Proof: Use definitions of $\mathbf{t}_k$ and $\bar{\mathbf{t}}_k$ and observe that the left hand side of (5.1.5) is the cross-product of these two vectors. Therefore the absolute value of the left hand side of (5.1.5) is equal to

$$\lVert \mathbf{t}_k(\sigma) \rVert \; \lVert \bar{\mathbf{t}}_k(\sigma) \rVert \; | \sin\angle(\mathbf{t}_k, \bar{\mathbf{t}}_k) | . \qquad \blacksquare$$

A variation of relation (5.2.1) is

$$\lVert \mathbf{t}_k(\sigma) \rVert \; \lVert \mathbf{t}_k(\sigma) - \bar{\mathbf{t}}_k(\sigma) \rVert \; | \sin\angle(\mathbf{t}_k, \mathbf{t}_k - \bar{\mathbf{t}}_k) | \;=\; \beta_{k+1} \, | \mathbf{y}_k^t F_k \bar{\mathbf{y}}_k |, \quad k = 1, \ldots, n-1. \tag{5.2.2}$$

The combination of relation (5.2.2) with **Corollary 1 of Lemma 4.2** that

$$\lVert \mathbf{t}_k \rVert \; \lVert \mathbf{t}_k' \rVert \; | \sin\angle(\mathbf{t}_k, \mathbf{t}_k') | \;=\; \beta_{k+1} \tag{5.2.3}$$

yields the following important relation:

Theorem 5.1: *For any real* σ *and* $k = 1, \ldots, n-1$

$$\lVert \mathbf{t}_k(\sigma) - \bar{\mathbf{t}}_k(\sigma) \rVert \; | \sin\angle(\mathbf{t}_k, \mathbf{t}_k - \bar{\mathbf{t}}_k) | \;=\; \lVert \mathbf{t}_k' \rVert \; | \sin\angle(\mathbf{t}_k, \mathbf{t}_k') | \, | \mathbf{y}_k^t F_k \bar{\mathbf{y}}_k | . \tag{5.2.4}$$

The direct consequences of **Theorem 5.1** are:

$$\lVert \mathbf{t}_k(\sigma) - \bar{\mathbf{t}}_k(\sigma) \rVert \;\ge\; \lVert \mathbf{t}_k' \rVert \; | \sin\angle(\mathbf{t}_k, \mathbf{t}_k') | \, | \mathbf{y}_k^t F_k \bar{\mathbf{y}}_k |, \qquad k = 1, \ldots, n-1, \tag{5.2.5}$$

$$= \frac{\beta_{k+1} \, | \mathbf{y}_k^t F_k \bar{\mathbf{y}}_k |}{\lVert \mathbf{t}_k \rVert} \tag{5.2.6}$$

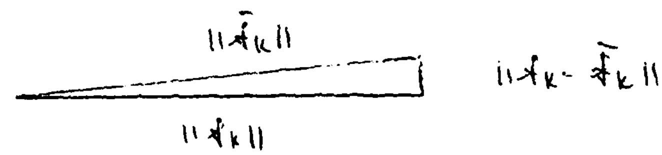
and

$$\lVert \mathbf{t}_k' \rVert \;\ge\; \frac{\lVert \mathbf{t}_k(\sigma) - \bar{\mathbf{t}}_k(\sigma) \rVert \; | \sin\angle(\mathbf{t}_k, \mathbf{t}_k - \bar{\mathbf{t}}_k) |}{| \mathbf{y}_k^t F_k \bar{\mathbf{y}}_k |}, \qquad k = 1, \ldots, n-1. \tag{5.2.7}$$

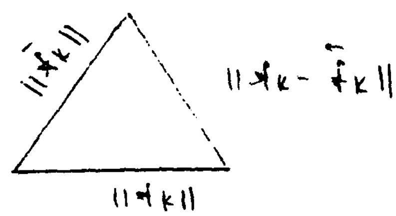
Now we come back to relation (5.2.1)

Since $|\,\beta_{k+1}\, \mathbf{y}_k^t F_k \bar{\mathbf{y}}_k\,|/2$ is the *AREA* of the triangular defined by $\mathbf{t}_k$, $\bar{\mathbf{t}}_k$ and $\mathbf{t}_k - \bar{\mathbf{t}}_k$. We shall describe instability in terms of this geometric figure. There are three stages:

a) stability



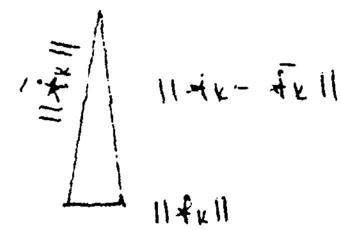
b) onset of instability



characterized by relative error in $\|\mathbf{t}_k\|$ such that

$$\frac{\|\mathbf{t}_k - \bar{\mathbf{t}}_k\|}{\|\mathbf{t}_k\|} = \frac{\|\mathbf{t}_k - \bar{\mathbf{t}}_k\|}{\|\bar{\mathbf{t}}_k\|} = 1.$$

c) full instability



We can expect to lose all correct digits in the greater of $|\pi_k|$ and $|c_k|\,\beta_{k+1}$ when the triangle O, $\mathbf{t}_k$, $\bar{\mathbf{t}}_k$ is equilateral:

$$\frac{\sqrt{3}}{2}\|\mathbf{t}_k\|^2 = \frac{\sqrt{3}}{2}\|\bar{\mathbf{t}}_k\|^2 = \beta_{k+1}\,|\,\mathbf{y}_k^t F_k \bar{\mathbf{y}}_k\,|.$$

Example 5.1:

The matrix is the same as that in **Example 2.2** with the same shift. The lengths of the residual vectors from *TQR* at step k are presented. For comparison, the correct lengths are also presented to show the relationship between the accuracy of $\|\bar{\mathbf{t}}_k\|$ and the magnitude of $\|\mathbf{t}_k\|$. It is not hard to observe that the product of corresponding elements in first column and third column is less than *spread* ε.

k	$\|\mathbf{t}_k(\lambda_6)\|$	$\|\bar{\mathbf{t}}_k(\lambda_6)\|$	$\|\mathbf{t}_k(\lambda_6) - \bar{\mathbf{t}}_k(\lambda_6)\|$
1	3.64965820*d*+00	3.64965820*d*+00	0.00000000*d*+00
2	3.16227602*d*−03	3.16227602*d*−03	4.15999586*d*−17
3	9.35882271*d*−07	9.35882271*d*−07	2.42128268*d*−13
4	2.37408198*d*−10	8.51726993*d*−10	8.18091259*d*−10
5	4.48883862*d*−14	3.22498160*d*−06	3.22498160*d*−06
6	1.60806628*d*−17	1.70563083*d*−02	1.70563083*d*−02

k	$\lvert \sin \angle(\mathbf{t}_k, \bar{\mathbf{t}}_k) \rvert$	$\lvert \sin \angle(\mathbf{t}_k, \bar{\mathbf{t}}_k - \mathbf{t}_k) \rvert$
1	0.00000d+00	1.00000d+00
2	5.26836d−09	1.00000d+00
3	2.58686d−07	1.00000d+00
4	9.60509d−01	1.00000d+00
5	1.00000d+00	1.00000d+00
6	0.00000d+00	0.00000d+00

Example 5.2:

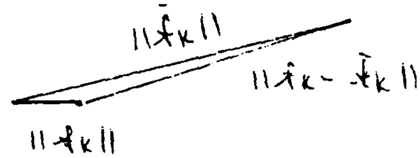
Another interesting example is one sweep of *TQR* on the following matrix:

$$T_{25} \;=\; \mathit{tridiagonal} \begin{bmatrix} & 1 & 1 & & 1 & 1 & \\ 13 & 0 & 0 \cdots & 0 & 0 & 13 \\ & 1 & 1 & & 1 & 1 & \end{bmatrix}.$$

The shift is its largest eigenvalue. We exhibit the $\|\mathbf{t}_k\|$, $\|\bar{\mathbf{t}}_k\|$, $\|\bar{\mathbf{t}}_k - \mathbf{t}_k\|$, $\lvert \sin \angle(\mathbf{t}_k, \bar{\mathbf{t}}_k) \rvert$ and $\lvert \sin \angle(\mathbf{t}_k, \bar{\mathbf{t}}_k - \mathbf{t}_k) \rvert$ in the following table.

1	1.0030d+00	1.0030d+00	0.0000d+00	0.00d+00	1.00d+00
2	7.6923d-02	7.6923d-02	0.0000d+00	0.00d+00	1.00d+00
3	5.9171d-03	5.9171d-03	3.0000d-14	0.00d+00	9.97d-01
4	4.5516d-04	4.5516d-04	3.7620d-13	0.00d+00	9.88d-01
5	3.5012d-05	3.5012d-05	4.8893d-12	1.38d-07	9.88d-01
6	2.6932d-06	2.6933d-06	6.3561d-11	2.33d-05	9.88d-01
7	2.0717d-07	2.0730d-07	8.2629d-10	3.94d-03	9.88d-01
8	1.5936d-08	2.0536d-08	1.0742d-08	5.17d-01	9.88d-01
9	1.2259d-09	1.3984d-07	1.3964d-07	9.87d-01	9.88d-01
10	9.4298d-11	1.8154d-06	1.8154d-06	9.88d-01	9.88d-01
11	7.2532d-12	2.3600d-05	2.3600d-05	9.88d-01	9.88d-01
12	5.5304d-13	3.0680d-04	3.0680d-04	9.97d-01	9.97d-01
13	5.5304d-13	3.9883d-03	3.9883d-03	7.67d-02	7.67d-02
14	7.2532d-12	5.1848d-02	5.1848d-02	4.50d-04	4.50d-04
15	9.4298d-11	6.7313d-01	6.7313d-01	2.66d-06	2.66d-06
16	1.2259d-09	7.2659d+00	7.2659d+00	1.67d-08	1.58d-08
17	1.5936d-08	1.2916d+01	1.2916d+01	0.00d+00	0.00d+00
18	2.0717d-07	1.2999d+01	1.2999d+01	0.00d+00	0.00d+00
19	2.6932d-06	1.3000d+01	1.3000d+01	5.27d-09	5.27d-09
20	3.5012d-05	1.3000d+01	1.3000d+01	5.27d-09	0.00d+00
21	4.5516d-04	1.3000d+01	1.3000d+01	0.00d+00	0.00d+00
22	5.9171d-03	1.3000d+01	1.3006d+01	0.00d+00	0.00d+00
23	7.6920d-02	1.3000d+01	1.3077d+01	0.00d+00	0.00d+00
24	9.9704d-01	1.3000d+01	1.3997d+01	0.00d+00	0.00d+00

The behavior of $\|\mathbf{t}_k\|$ has gone through the three stages described before and the instability occurs at *step* 8. However starting from *step* 13, the exact $\|\mathbf{t}_k\|$ begins to increase and the computed $\|\bar{\mathbf{t}}_k\|$ still keeps increasing. As a result, the angle between these two vectors has to shrink in order to satisfy the relation (5.2.1). Finally the triangular turns into the following form:



The interesting point of this example is that the computed $\|\bar{\mathbf{t}}_{24}\|$ can be totally wrong even though exact $\|\mathbf{t}_{24}\|$ is not tiny. However, the instability occurred when $\|\bar{\mathbf{t}}_8\|$ was about $O(\mathit{spread}\,\varepsilon)$.

Explanation of forward Instability

The **Corollary of Lemma 5.1**, coupled with the behavior of $\mathbf{y}_k^t F_k \bar{\mathbf{y}}_k$, shows why instability must occur in certain cases.

There are classes of tridiagonal matrices and shift σ such that $\|\mathbf{t}_k\|$ can diminish with k to as small a value as desired. Furthermore for small enough k the vectors $\mathbf{y}_k$ and $\bar{\mathbf{y}}_k$ are nearly identical. As k increases the last components c_k and $\bar{c}_k$ may decrease to $O(\sqrt{\varepsilon})$, at which size $\bar{c}_k$ could lose all its significant digits. Nevertheless $\mathbf{y}_k$ is an unit vector and $\mathbf{y}_k^t \bar{\mathbf{y}}_k \approx 1$ (i.e. > 0.8). At this stage $\mathbf{y}_k^t F_k \bar{\mathbf{y}}_k$ is close to the Rayleigh quotient of F_k for $\mathbf{y}_k$, namely $\mathbf{y}_k^t F_k \mathbf{y}_k$. For typical roundoff situations this Rayleigh quotient will be of the same order as $\|F_k\|$. Recall from **Corollary of Lemma 5.2** that $\|F_k\| \le 5.6\,\mathit{spread}(T_k)\,\varepsilon$ when $\lambda_{\min}(T_k) \le \sigma \le \lambda_{\max}(T_k)$.

CONCLUSION:

a) If $|\sin\angle(\mathbf{t}_k, \bar{\mathbf{t}}_k - \mathbf{t}_k)| \ge \sqrt{3}/2$ and $\|\mathbf{t}_k\| \ge 4(\beta_{k+1}\,\mathit{spread}\,\varepsilon)^{1/2}$ then

$$\frac{\|\bar{\mathbf{t}}_k - \mathbf{t}_k\|}{\|\mathbf{t}_k\|} = \frac{\beta_{k+1}\,|\mathbf{y}_k^t F_k \bar{\mathbf{y}}_k|}{\|\mathbf{t}_k\|^2\,|\sin\angle(\mathbf{t}_k, \bar{\mathbf{t}}_k - \mathbf{t}_k)|} \le \frac{5.6\,\beta_{k+1}\,\mathit{spread}\,\varepsilon}{16\,\beta_{k+1}\,\mathit{spread}\,\varepsilon}\,\frac{2}{\sqrt{3}} < \frac{1}{2}.$$

b) While $|\mathbf{y}_k^t F_k \bar{\mathbf{y}}_k| > \frac{1}{4}\mathit{spread}\,\varepsilon$, we have

$$\frac{\|\bar{\mathbf{t}}_k - \mathbf{t}_k\|}{\|\mathbf{t}_k\|} = \frac{\beta_{k+1}\,|\mathbf{y}_k^t F_k \bar{\mathbf{y}}_k|}{\|\mathbf{t}_k\|^2\,|\sin\angle(\mathbf{t}_k, \bar{\mathbf{t}}_k - \mathbf{t}_k)|} \ge \frac{\beta_{k+1}\,\mathit{spread}\,\varepsilon}{4\,\|\mathbf{t}_k\|^2}.$$

Thus as $\|\mathbf{t}_k\|^2$ drops below $(\beta_{k+1}\,\mathit{spread}\,\varepsilon)/4$ so must the relative error in $\bar{\mathbf{t}}_k$ rise drastically above 1. This is complete instability.

Note: we know of one special case (see [Demmel and Kahan]) when all diagonal elements of T are zero and a variation of *TQR* with shift as zero forces this condition on the diagonal elements of the transformed T. In this case it turns out that $\mathbf{y}_k^t F_k \bar{\mathbf{y}}_k$ is structurally zero for all k. Thus we can not prove that $\mathbf{y}_k^t F_k \bar{\mathbf{y}}_k$ in general remains above $\mathit{spread}\,\varepsilon/4$ until complete instability ($\mathbf{y}_k^t \bar{\mathbf{y}}_k \approx O(\sqrt{\varepsilon})$) sets in. However this is what we observe.

5.3 Floating point version of TQR

In the following, we want to present a floating point version of the useful relation (3.3.8):

$$(T_k - \sigma I_k)\,\mathbf{y}_k = \pi_k\,\mathbf{e}_k^{(k)}$$

Our model of error in floating point arithmetic is

$$fl(x \circ y) = (x \circ y)(1 + e) \tag{5.3.1}$$

where o is one of +, −, × and /; $fl(x \circ y)$ is the floating point result of the operation o, and $|e| \le \varepsilon$, where ε is the machine precision. Our analysis will be linearized using *Wilkinson*'s notation (see [Wilkinson, pp 113]):

If $|e_i| \le \varepsilon$, $i = 1, \ldots, n$ *and* $n\varepsilon \le 0.01$, *then*

$$\prod_{i=1}^{n} (1 + e_i) = 1 + 1.01\,\theta\, n\, \varepsilon \tag{5.3.2}$$

where $|\theta| \le 1$.

Lemma 5.2: *Let c_i, s_i and π_i denote the outputs of* **TQR** *in exact arithmetic with shift σ. Let $\bar{c}_i$, $\bar{s}_i$ and $\bar{\pi}_i$ denote the outputs of* **TQR** *in finite precision arithmetic with shift σ. Then the computed outputs from* **TQR** *with shift σ satisfy exactly the following relation:*

$$(T_k - \sigma I_k + F_k)\bar{y}_k = \bar{\pi}_k e_k^{(k)} \tag{5.3.3}$$

where

$$F_k = \begin{bmatrix} (\alpha_1-\sigma)\varepsilon_{1,1} & \beta_2\varepsilon_{1,2} & & & \\ \beta_2\varepsilon_{2,1} & (\alpha_2-\sigma)\varepsilon_{2,2} & \beta_3\varepsilon_{2,3} & & \\ & \beta_3\varepsilon_{3,2} & \cdot & \cdot & \\ & & \cdot & \cdot & \beta_k\varepsilon_{k-1,k} \\ & & & \beta_k\varepsilon_{k,k-1} & (\alpha_k-\sigma)\varepsilon_{k,k} \end{bmatrix}$$

and

$$|\varepsilon_{k,k}| \le 3.03\varepsilon, \quad |\varepsilon_{k,k-1}| \le 3.03\varepsilon, \quad |\varepsilon_{k,k+1}| \le 2.02\varepsilon, \quad \text{for } k = 1, 2, \ldots, n.$$

Proof: We use mathematical induction.

Use definition of π_1 in (4.1.1) to find $\pi_1 = \alpha_1 - \sigma$. Hence by (5.3.1)

$$\bar{\pi}_1 = fl(\pi_1) = fl(\alpha_1 - \sigma) = (\alpha_1 - \sigma)(1 + \varepsilon_{1,1}). \tag{5.3.4}$$

Use definition of π_k in (4.1.1) and then (5.3.1), (5.3.2) to find

$$\begin{aligned} \bar{\pi}_k &= fl(\beta_k \bar{c}_{k-1}(-\bar{s}_k) + (\alpha_k - \sigma)\bar{c}_k) \\ &= \beta_k(1 + \varepsilon_{k,k-1})\bar{c}_{k-1}(-\bar{s}_k) + (\alpha_k - \sigma)(1 + \varepsilon_{k,k})\bar{c}_k, \end{aligned} \tag{5.3.5}$$

with $|\varepsilon_{k,k-1}| \le 3.03\varepsilon$, $|\varepsilon_{k,k}| \le 3.03\varepsilon$ because each quantity is involved in 3 atomic arithmetic operations. Do the same thing for c_k, s_k,

$$\bar{c}_k = fl(\bar{\pi}_{k-1}/\bar{\xi}_k) = \frac{\bar{\pi}_{k-1}}{\bar{\xi}_k(1 + \varepsilon_{ck})}, \qquad \bar{s}_k = fl(\beta_k/\bar{\xi}_k) = \frac{\beta_k(1 + \varepsilon_{sk})}{\bar{\xi}_k}.$$

Then

$$\bar{s}_k \bar{\pi}_{k-1} = \beta_k \bar{c}_k (1 + \varepsilon_{ck})(1 + \varepsilon_{sk}) = \beta_k \bar{c}_k (1 + \varepsilon_{k-1,k}), \tag{5.3.6}$$

with $|\varepsilon_{k-1,k}| \le 2.02\varepsilon$.

Now we start to put things into matrix form. Consider the case $k = 2$. Multiplying (5.3.4) by $\bar{s}_2$,

$$\bar{s}_2 \bar{\pi}_1 = (\alpha_1 - \sigma)(1 + \varepsilon_{1,1})\bar{c}_1 \bar{s}_2 \qquad \textit{since} \quad \bar{c}_1 = 1. \tag{5.3.7}$$

Use (5.3.6) and (5.3.5) for $k = 2$,

$$\bar{s}_2\bar{\pi}_1 = \beta_2(1+\varepsilon_{1,2})\bar{c}_2 = (\alpha_1-\sigma)(1+\varepsilon_{1,1})\bar{c}_1\bar{s}_2; \qquad \text{by (5.3.7)}$$
$$\bar{\pi}_2 = \beta_2(1+\varepsilon_{2,1})\bar{c}_1(-\bar{s}_2) + (\alpha_2-\sigma)(1+\varepsilon_{2,2})\bar{c}_2.$$

That is to say

$$\begin{bmatrix} (\alpha_1-\sigma)(1+\varepsilon_{1,1}) & \beta_2(1+\varepsilon_{1,2}) \\ \beta_2(1+\varepsilon_{2,1}) & (\alpha_2-\sigma)(1+\varepsilon_{2,2}) \end{bmatrix} \begin{bmatrix} \bar{c}_1(-\bar{s}_2) \\ \bar{c}_2 \end{bmatrix} = \begin{bmatrix} 0 \\ \bar{\pi}_2 \end{bmatrix}. \tag{5.3.8}$$

The formulae (5.3.4) and (5.3.8) have shown that (5.3.3) is true for $k = 1, 2$. Now suppose the statement is true for $k = j$. That is to say

$$\begin{bmatrix} (\alpha_1-\sigma)(1+\varepsilon_{1,1}) & \beta_2(1+\varepsilon_{1,2}) & & & \\ \beta_2(1+\varepsilon_{2,1}) & (\alpha_2-\sigma)(1+\varepsilon_{2,2}) & \beta_3(1+\varepsilon_{2,3}) & & \\ & \beta_3(1+\varepsilon_{3,2}) & \cdot & \cdot & \\ & & \cdot & \cdot & \beta_j(1+\varepsilon_{j-1,j}) \\ & & & \beta_j(1+\varepsilon_{j,j-1}) & (\alpha_j-\sigma)(1+\varepsilon_{j,j}) \end{bmatrix}$$
$$\times \begin{bmatrix} \bar{c}_1(-\bar{s}_2)\cdots(-\bar{s}_j) \\ \bar{c}_2(-\bar{s}_3)\cdots(-\bar{s}_j) \\ \cdot \\ \cdot \\ \bar{c}_j \end{bmatrix} = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ 0 \\ \bar{\pi}_j \end{bmatrix}. \tag{5.3.9}$$

Multiply by $(-\bar{s}_{j+1})$ on both sides of the above equation and use the results (5.3.6), (5.3.5) for $k = j+1$, namely

$$(-\bar{s}_{j+1})\bar{\pi}_j = -\beta_{j+1}(1+\varepsilon_{j,j+1})\bar{c}_{j+1},$$
$$\beta_{j+1}(1+\varepsilon_{j+1,j})\bar{c}_j(-\bar{s}_{j+1}) + (\alpha_{j+1}-\sigma)(1+\varepsilon_{j+1,j+1})\bar{c}_{j+1} = \bar{\pi}_{j+1}.$$

Now adjust the last row of (5.3.9) and add a new row to get

$$\begin{bmatrix} (\alpha_1-\sigma)(1+\varepsilon_{1,1}) & \beta_2(1+\varepsilon_{1,2}) & & & & \\ \beta_2(1+\varepsilon_{2,1}) & (\alpha_2-\sigma)(1+\varepsilon_{2,2}) & \beta_3(1+\varepsilon_{2,3}) & & & \\ & \beta_3(1+\varepsilon_{3,2}) & \cdot & \cdot & & \\ & & \cdot & \cdot & \beta_j(1+\varepsilon_{j-1,j}) & \\ & & & \beta_j(1+\varepsilon_{j,j-1}) & (\alpha_j-\sigma)(1+\varepsilon_{j,j}) & \beta_{j+1}(1+\varepsilon_{j,j+1}) \\ & & & & \beta_{j+1}(1+\varepsilon_{j+1,j}) & (\alpha_{j+1}-\sigma)(1+\varepsilon_{j+1,j+1}) \end{bmatrix}$$
$$\times \begin{bmatrix} \bar{c}_1(-\bar{s}_2)\cdots(-\bar{s}_{j+1}) \\ \bar{c}_2(-\bar{s}_3)\cdots(-\bar{s}_{j+1}) \\ \cdot \\ \cdot \\ \bar{c}_j(-\bar{s}_{j+1}) \\ \bar{c}_{j+1} \end{bmatrix} = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ \bar{\pi}_{j+1} \end{bmatrix}.$$

By induction, (5.3.3) holds for $k \le n$. ■

Corollary of Lemma 5.2 *If*

$$\lambda_{\min}(T_n) \le \sigma \le \lambda_{\max}(T_n)$$

then

$$\| F_n \| < 5.6\, spread(T_n)\, \varepsilon. \tag{5.3.10}$$

where $spread(T_n) = \lambda_{\max}(T_n) - \lambda_{\min}(T_n)$.

Proof:

(1) by **Lemma 5.1**

$|\varepsilon_{k,k}| \le 3.03\varepsilon$, $|\varepsilon_{k,k-1}| \le 3.03\varepsilon$, $|\varepsilon_{k,k+1}| \le 2.02\varepsilon$, for $k = 1, 2, ..., n$;

(2) $|\beta_k| \le spread(T_n)/2$ for $2 \le k \le n$, by **Theorem 2.2**;

(3) $|\alpha_k - \sigma| \le spread(T_n)$ for $1 \le k \le n$, by condition given,

then, abbreviating $spread(T_n)$ by *spread*,

$$\begin{aligned} \| F_n \|_\infty &= \max_k [\, |\beta_k \varepsilon_{k,k-1}| + |(\alpha_k - \sigma)\varepsilon_{k,k}| + |\beta_{k+1}\varepsilon_{k,k+1}| \,] \\ &\le \frac{3}{2}\, spread\ 1.01\varepsilon + 3\, spread\ 1.01\varepsilon + \frac{2}{2}\, spread\ 1.01\varepsilon \\ &= (3/2 + 3 + 1)\ 1.01\, spread\ \varepsilon \\ &< 5.6\, spread\ \varepsilon. \end{aligned}$$

In the same way,

$$\| F_n \|_1 < 5.6\, spread\ \varepsilon.$$

Hence

$$\begin{aligned} \| F_n \|^2 &= \max \lambda((F_n)^t F_n) \\ &\le \| (F_n)^t F_n \|_\infty \\ &\le \| (F_n)^t \|_\infty \| F_n \|_\infty \\ &= \| F_n \|_1 \| F_n \|_\infty \\ &< (5.6\, spread\ \varepsilon)^2. \qquad \blacksquare \end{aligned}$$

Note: When σ is outside the interval $[\lambda_{\min}(T_n), \lambda_{\max}(T_n)]$ we can bound term $|\alpha_k - \sigma|$ by $\mu_n + spread(T_n)$. Therefore for any real σ

$$\| F_n \| < 5.6\, spread(T_n)\varepsilon + 3.03\mu_n\varepsilon. \tag{5.3.11}$$

5.4 Forward stability

In this subsection, we will present a forward perturbation analysis. It turns out that the instability can occur only if σ is very close to an eigenvalue of T_k.

Definition 5.2:

$$\hat{y}_k := \bar{y}_k / \|\bar{y}_k\|, \qquad \hat{\pi}_k := \bar{\pi}_k / \|\bar{y}_k\|.$$

Recall that $\mu_k(\sigma) = \min_i |\sigma - \lambda_i^{(k)}|$ and $spread(T_k) = \lambda_{\max}(T_k) - \lambda_{\min}(T_k)$.

Theorem 5.2 *For any real* σ, *if*

$$\mu_k(\sigma) > 12\, spread(T_k)\varepsilon, \qquad k = 1, ..., n \tag{5.4.1}$$

then

$$\varepsilon_{\pi(k)} := \frac{|\hat{\pi}_k - \pi_k|}{|\pi_k|} < \frac{1}{2}, \qquad k = 1, \ldots, n. \tag{5.4.2}$$

Proof: Let $B_k = (T_k - \sigma I_k)^{-1} F_k$. Use (5.1.3) to find

$$\begin{aligned}\hat{\mathbf{y}}_k &= (T_k - \sigma I_k + F_k)^{-1} \mathbf{e}_k \hat{\pi}_k, \\ &= (I + B_k)^{-1} (T_k - \sigma I_k)^{-1} \mathbf{e}_k \hat{\pi}_k, \\ &= (I + B_k)^{-1} \mathbf{y}_k \hat{\pi}_k / \pi_k. \end{aligned} \tag{5.4.3}$$

Premultiply by $\mathbf{y}_k^t$ to get the main formula

$$\mathbf{y}_k^t \hat{\mathbf{y}}_k = \mathbf{y}_k^t (I + B_k)^{-1} \mathbf{y}_k \hat{\pi}_k / \pi_k. \tag{5.4.4}$$

(1) we show

$$\| B_k \| \le 1/2, \qquad k = 1, 2, \ldots, n. \tag{5.4.5}$$

$$\begin{aligned}\|B_k\| &\le \| (T_k - \sigma I_k)^{-1} \| \, \|F_k\| \le \frac{\|F_k\|}{\mu_k}, \\ &\le \frac{3.03\,\mu_k \varepsilon + 5.6\, spread(T_k)\, \varepsilon}{\mu_k}, && \text{by (5.3.11)} \\ &< 1/2, && \text{by (5.4.1).} \ \blacksquare \end{aligned}$$

(2) we show

$$\mathbf{y}_k^t (I + B_k)^{-1} \mathbf{y}_k > 0, \qquad k = 1, 2, \ldots, n. \tag{5.4.6}$$

$$\begin{aligned}\mathbf{y}_k^t (I + B_k)^{-1} \mathbf{y}_k &= 1 + \sum_{i=1}^{\infty} (-1)^i \mathbf{y}_k^t B_k^i \mathbf{y}_k \\ &\ge 1 - \sum_{i=1}^{\infty} \|B_k^i\|, \\ &\ge 1 - \frac{\|B_k\|}{1 - \|B_k\|}, \\ &= \frac{1 - 2\|B_k\|}{1 - \|B_k\|}, \\ &> 0, && \text{by (5.4.5).} \end{aligned}$$

(3) we show by induction that

$$\mathbf{y}_k^t \hat{\mathbf{y}}_k > 0, \qquad \pi_k \hat{\pi}_k > 0, \qquad k = 1, 2, \ldots, n. \tag{5.4.7}$$

Note that $\mathbf{y}_1^t \hat{\mathbf{y}}_1 = 1 > 0$ and $\pi_1 \hat{\pi}_1 > 0$ by (5.4.4) and (5.4.6) with $k = 1$. Now suppose that

$$\mathbf{y}_j^t \hat{\mathbf{y}}_j > 0, \qquad \pi_j \hat{\pi}_j > 0, \tag{5.4.8}$$

then

$$c_{j+1} \bar{c}_{j+1} = \frac{\pi_j}{\xi_{j+1}} \frac{\bar{\pi}_j}{\bar{\xi}_{j+1}} (1 + \varepsilon) = \frac{\pi_j}{\xi_{j+1}} \frac{\hat{\pi}_j \|\bar{\mathbf{y}}_j\|}{\bar{\xi}_{j+1}} (1 + \varepsilon) > 0. \tag{5.4.9}$$

Reference to (5.1.1) shows that

$$\mathbf{y}_{j+1}^t \hat{\mathbf{y}}_{j+1} = \mathbf{y}_{j+1}^t \bar{\mathbf{y}}_{j+1} / \|\bar{\mathbf{y}}_{j+1}\| = \frac{1}{\|\bar{\mathbf{y}}_{j+1}\|} \left(\mathbf{y}_j^t \hat{\mathbf{y}}_j \|\bar{\mathbf{y}}_j\| s_{j+1} \bar{s}_{j+1} + c_{j+1} \bar{c}_{j+1} \right).$$

Since $s_{j+1} > 0$, $\bar{s}_{j+1} > 0$ (because $\beta_{j+1} > 0$) then by (5.4.8) and (5.4.9) we see that

$$\mathbf{y}_{j+1}^t \hat{\mathbf{y}}_{j+1} > 0.$$

Use (5.4.4) and (5.4.6) with $k = j+1$ to find $\pi_{j+1} \hat{\pi}_{j+1} > 0$. By principle of induction, (5.4.7) is true

for $k = 1, \ldots, n$.

(4) we show (5.4.2).

Take norms of (5.4.3) and use (5.4.7) to get:

$$\pi_k/\hat{\pi}_k \;=\; \| (I + B_k)^{-1} \mathbf{y}_k \|$$

Use standard perturbation theory and (5.4.5) to find:

$$\| (I + B_k)^{-1} \mathbf{y}_k \| \;\le\; \| (I + B_k)^{-1} \| \;\le\; \frac{1}{1 - \|B_k\|}.$$

Let $\mathbf{y}_k = (I + B_k)\mathbf{w}$. Then $1 \le (1 + \|B_k\|)\,\|\mathbf{w}\|$. Hence

$$\frac{1}{1 + \|B_k\|} \;\le\; \| (I + B_k)^{-1} \mathbf{y}_k \|.$$

Therefore

$$\frac{1}{1 + \|B_k\|} \;\le\; \pi_k/\hat{\pi}_k \;\le\; \frac{1}{1 - \|B_k\|}.$$

So

$$\varepsilon_{\pi(k)} \;=\; \frac{|\hat{\pi}_k - \pi_k|}{|\pi_k|} \;\le\; \|B_k\| \;<\; \frac{1}{2}. \qquad \blacksquare$$

Conclusion

Forward instability was introduced as the occurrence of a computed transform $\hat{T}$ that was far from the exact $\hat{T}$. However our study concentrates attention on the forward stability of the rotation angles that determine the similarity transformation. This is reasonable because wildly erroneous angles ensures both a wrong $\hat{T}$ and also a vector $\bar{\mathbf{y}}_n$ that is not a reasonable eigenvector approximation. However even rotation angles with high relative accuracy cannot guarantee that each computed element of $\hat{T}$ has high relative accuracy.

In **section 5.4** we have shown that the computed π_k will not lose all the significant digits if the shift σ is nowhere close to any eigenvalues of T_j, $j = 1, \ldots, k$. That is equivalent to say that instability can only happen at *step* k if σ has been very close to some eigenvalue of T_j, $j < k$. However σ being close to an eigenvalue itself does not provoke forward instability. **Example 2.4** illustrated this fact.

We point out some main results of this study.

(1) The effect on c_k and s_k of changes in the matrix elements α_i and β_i is transmitted through a perturbation matrix F_k. By **Theorem 5.1**

$$\|\bar{\mathbf{t}}_k - \mathbf{t}_k\| \, |\sin \angle(\mathbf{t}_k, \bar{\mathbf{t}}_k - \mathbf{t}_k)| \;<\; \|\mathbf{t}_k'\| \, |\mathbf{y}_k^t F_k \bar{\mathbf{y}}_k|.$$

and the amplification factor for the $|\mathbf{y}_k^t F_k \bar{\mathbf{y}}_k|$ is $\|\mathbf{t}_k'\|$, the derivative with respect to σ. Instability requires that $\|\mathbf{t}_k'\|$ be huge. There is no need to compute the gradient of $\|\mathbf{t}_k\|$ with respect to all the variables α_i and β_i.

(2) Formula (5.2.6) shows that

$$\frac{\|\mathbf{t}_k - \bar{\mathbf{t}}_k\|}{\|\mathbf{t}_k\|} \;\ge\; \frac{\beta_{k+1} \, |\mathbf{y}_k^t F_k \bar{\mathbf{y}}_k|}{\|\mathbf{t}_k\|^2}.$$

Thus whenever $\|\mathbf{t}_k\|$ gets too small, the relative error in $\|\mathbf{t}_k\|$ can rise well above 1. In other words small enough values of $|c_k|$ and $|\pi_k|$ cannot be calculated accurately in finite

precision arithmetic. Actually when $\|t_k\|$ is tiny, π_k^2 and $c_k^2\beta_{k+1}^2$ will be tiny. Therefore current shift σ has to be very close to an eigenvalue of T_k. Then by (4.3.3) and (4.1.6) we see that $1/|c_k\beta_{k+1}|$ will be very close to the derivative of c_{k+1} at this σ. Therefore tiny $\|t_k\|$ signals huge derivative on c_{k+1}. Using **Theorem 4.4**, we see that μ_k and μ_{k+1} have to be tiny simultaneously.

Bibliography

1. Cao, Z. H., "The Matrix Eigenvalue Problem", *Shanghai Science-Technology Publishing Company, Shanghai*, **1980**.
2. Cauchy, A., Cours D'Analyse de L'Ecole Polytechnique *Oeuvres Completes, vol. 2 and 3*, **1821**.
3. Demmel, J.; Kahan, W., Computing Small Singular Values of Bidiagonal Matrices With Guaranteed High Relative Accuracy, **1987**, To appear.
4. Golub, G. H.; Van Loan, C. F., "Matrix Computation", *The Johns Hopkins University Press, Baltimore, Maryland*, **1983**.
5. Golub, G. H.; Kahan, W., Calculating the Singular Values and Pseudo-inverse of a Matrix, *SIAM J. Num. Anal.*, **1965**, *vol. 2*, pp. 205-224.
6. Parlett, B. N., "The Symmetric Eigenvalue Problem", *Prentice-Hall, Englewood Cliffs, N.J.*, **1980**.
7. Smith, B. T.; Boyle, J. M.; Garbow, B. S.; Ikebe, Y.; Klema, V. C.; Moler, C. B., Matrix Eigensystem Routines - EISPACK Guide, "Lecture Notes in Computer Science, Second Edition, vol.6", *Springer, New York*, **1976**.
8. Stewart, G. W., "Introduction to Matrix Computation", *Academic Press, New York*, **1973**.
9. Stewart, G. W., Incorporating Origin Shifts into the QR Algorithm for Symmetric Tridiagonal Matrices, *Comm. Assoc. Comp. Mach.*, **1970**, *vol. 13*, pp. 365-367.
10. Wilkinson, J. H., "The Algebraic Eigenvalue Problem", *Oxford University Press, London*, **1965**.
11. Wilkinson, J. H., The calculation of the eigenvectors of codiagonal matrices, *Comput. J.*, **1958**, *vol. 1*, pp. 90-96.

cccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccc

the eigenvalues of the leading submatrices of T(6)

cccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccc

T(1)'s	6.6833333333d-01					
T(2)'s	1.9989994996d-03	3.9999974988d+00				
T(3)'s	5.8543640493d-04	3.4138632448d-03	4.0000000000d+00			
T(4)'s	1.5598522605d-04	1.9996973277d-03	3.8438029033d-03	4.0000000000d+00		
T(5)'s	2.4815314197d-05	1.2737235803d-03	2.7260330323d-03	3.9751422160d-03	4.0000000000d+00	
T(6)'s	9.5335916983d-17	1.0000000000d-03	2.0000000000d-03	3.0000000000d-03	4.0000000000d-03	4.0000000000d+00

table 2.3.1

cccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccc

the eigenvalues of the leading submatrices of T(6) after one QR with shift=4

cccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccc

T(1)'s	1.9989994999d-03					
T(2)'s	5.8543640493d-04	3.4138632448d-03				
T(3)'s	1.5598522605d-04	1.9996973277d-03	3.8438029033d-03			
T(4)'s	2.4815314390d-05	1.2737235814d-03	2.7260330334d-03	3.9751422162d-03		
T(5)'s	9.9691264906d-07	1.0158927266d-03	2.0373741571d-03	3.0174162845d-03	4.0010857232d-03	
T(6)'s	-3.4016207253d-18	1.0000000000d-03	2.0000000000d-03	3.0000000000d-03	4.0000000000d-03	4.0000000000d+00

table 2.3.2

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) University of California, Berkeley	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

On the Forward Instability of the QR Transformation

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

CPAM report, July 1988

5. AUTHOR(S) (First name, middle initial, last name)

J. Le, B. N. Parlett

6. REPORT DATE	7a. TOTAL NO OF PAGES	7b. NO. OF REFS
July 1988	37	

8a. CONTRACT OR GRANT NO. N00014-76-C-0013	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO. c. d.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. DISTRIBUTION STATEMENT

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Mathematics Branch Office of Naval Research Washington, DC 20360

13. ABSTRACT

QR is the standard method for finding all the eigenvalues of a symmetric tridiagonal matrix. It produces a sequence of similar tridiagonals. It is well known that the QR transformation from T to $\hat{T}$ is backward stable. That means that the computed $\hat{T}$ is exactly orthogonally similar to a matrix close to T. It is also known that the algorithm sometimes exhibits forward instability. That means that the computed $\hat{T}$ is not close to the exact $\hat{T}$.

For the purpose of computing eigenvalues the property of backward stability is all that one requires. However the QR transformation has other uses and there forward stability is wanted.

This report analyzes the forward instability and shows that it occurs only when the shift causes premature deflation. We show that forward stability is governed by the behavior, in exact arithmetic, of a pair of variables and we establish tight upper and lower bounds on their derivatives with respect to change in the shift parameter.

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6801

Security Classification

END
Filmed
5-89
DTIC